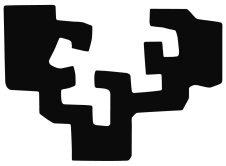
eman ta zabal zazu



EUSKAL HERRIKO UNIBERTSITATEA
Hizkuntzaren Azterketa eta Prozesamendua doktoretza-programa

Doktoretza-tesia

Leveraging Feedback in Conversational Question Answering Systems

Jon Ander Campos Tejedor

2023

EUSKAL HERRIKO UNIBERTSITATEA
Hizkuntzaren Azterketa eta Prozesamendua doktoretza-programa

Leveraging Feedback in Conversational Question Answering Systems

Jon Ander Campos Tejedorrek Eneko Agirre eta Gorka Azkuneren zuzendaritzapean eginiko tesi-txostena, Euskal Herriko Unibertsitatean Doktore titulua eskuratzeko aurkeztua.

Donostia, 2023ko Ekaina.

. . .

pasioa da hemen exijitzea zilegi den gutxieneko hori

. . .

Gorka Urbizu ("Poligrafo bakarra", Berri Txarrak)

. . .

When you depart for Ithaca,
wish for the road to be long,
full of adventure, full of knowledge.
Fear not the Laestrygonians and the Cyclopes,
nor the angry Poseidon.

. . .

Konstantinos P. Kavafis ("Ithaca")

Esker ona

Eskerrik asko...

... Aitor, Arantxa, Eneko eta Gorka emandako laguntza guztiagatik. Zuek erakutsi didazue ikertzaile izatea zer den eta zuek gabe ez nintzateke gaur naizena izango.

... Ixa taldeari eta Ixakideei ikerketa talde bat baina askoz gehiago ere izateagatik. Partekatutako bizipen guztiengatik, mikrouhin-labe bahituetatik krispeta festa klandestinoetara. Mila esker Josebari, garagardo, txapeldunen merienda eta surf sesio guztiengatik.

... Kyunghyun for giving me the chance to collaborate with you and hosting me at NYU. My work would not be the same without the wonderful collaborators I met during my visits: Angelica Chen, Ethan Perez, Jérémy Scheurer, Jun Shern Chan, Sam Bowman and Tomek Korbak. Thank you to Richard, Nitish and Vishakh for the enriching talks and dinners.

... Sophie and Sahar for hosting me at CNRS when I was just starting my PhD. It was a pleasure to collaborate with you during the early stages of my thesis.

... Satwik Kottur and Peter Cahill for giving me the chance to perform summer intenships at your teams at Meta and Apple. I learnt so much from you during these experiences.

... aita, ama eta Maialen uneoro hor egoteagatik. Zuek gabe tesi hau ez litzateke posible izango.

... Mona, Kris and especially Jordana. I never thought I would find a home so far from Urretxu.

... lagun guztioi eta bereziki Eyak Jodiuri. Azken urteotan buelta dexente eman ditut munduan zehar eta beti hor egon zarete.

Lan hau Ekonomia eta Lehiakortasun Ministerioaren diru-laguntza bati esker egin da (FPU18/01271).

Abstract

Language is the most common way to interact between humans and it is increasingly being used to interact with machines too. The increasing popularity of language as the tool for human computer interaction has been driven by the latest advances in natural language understanding (NLU) and generation (NLG).

The goal of this thesis is to exploit the interaction that deployed systems have with humans, leveraging human feedback as a learning and adaptation signal. We specially focus on lifelong adaptation and the domain shift that conversational systems can face when being deployed. For that purpose, our approach focuses on explicit binary feedback (correct, incorrect) as in most applications users are not able to provide the correct answer to the system.

To achieve our objective, we have first built a conversational question answering (CQA) dataset named DoQA, comprising 2,437 dialogues and 10,917 QA pairs collected from three Stack Exchange sites using the Wizard of Oz method with crowdsourcing. Compared to previous work, DoQA comprises well-defined information needs, leading to more coherent and natural conversations with less factoid questions and is multi-domain. We perform supervised experiments on top of this dataset in order to show the transfer learning capabilities of state-of-the-art systems. Furthermore, we empirically show the importance of modeling the conversational context for solving DoQA. This highlights the challenges that CQA datasets bring compared to traditional question answering (QA) datasets.

Having created DoQA, we propose feedback-weighted learning based on importance sampling to improve upon an initial supervised system using binary user feedback. We perform simulated experiments on CQA datasets where binary user feedback is derived from gold annotations. Our results demonstrate that our method is able to improve over an initial supervised system, achieving performance close to a fully-supervised system that has access to the same labeled examples in in-domain experiments and even out-of-domain experiments.

However, feedback-weighted learning has limitations when facing real world noisy user feedback. To address this issue, we propose a variation of the method that models the noise coming from users in the CQA task. Our negative results in this setting highlight the challenges in modeling the noise coming from users.

The contributions made in this thesis, along with parallel developments, have been utilized by the CQA community to advance the challenging task of domain adaptation and leveraging binary user feedback.

Laburpena

Hizkuntza gizakien artean komunikatzeko modurik ohikoena da eta makinekin komunikatzeko geroz eta gehiago erabiltzen ari da. Gizaki-ordenagailu elkarrekintzan hizkuntza tresna bezala irabazten ari den ospea hizkuntza naturalaren ulermen eta sorkuntzan izan diren azken aurrerapenek motibatu dute.

Tesi honen helburua, martxan jarri eta geroko sistemek gizakiekin duten elkarrekintza ustiatzea da, gizaki *feedback* bitarra ikasketa eta egokitzapen seinale bezala erabiliz. Fokua elkarrizketa sistemek martxan jartzerakoan jasaten duten domeinu aldaketan jartzen dugu. Helburu honetarako gure metodoan *feedback* bitar esplizitua erabiltzen dugu (zuzen, oker) kasu askotan erabiltzailea ez baita gai sistemari erantzun zuzena emateko.

Helburu hau lortzeko, lehenik eta behin DoQA izeneko galdera-erantzun motako elkarrizketez osatutako datu multzo bat sortzen dugu. Datu multzo honek 2.437 elkarrizketa eta 10.917 galdera-erantzun pare ditu *crowdworker*-ek *StackExchange*-ko hiru orrietatik *Wizard of Oz* metodoa jarraituz sortuak izan direnak. Aurretiko lanarekin konparatuta, DoQAk elkarrizketa koherente eta naturalagoak ditu, galdera konplexuez osatua dago eta domeinu anitzetako elkarrizketak ditu. Artearen egoerako sistemen transferentzia gaitasunak aztertzeko datu multzo honen gainean esperimentu gainbegiratuak egin ditugu. Honetaz gain, modu enpirikoan erakutsi dugu elkarrizketaren testuingurua kontuan hartzearen garrantzia DoQA ebazteko garaian.

DoQA sortu ostean, *feedback* bitarretik ikasteko algoritmo bat definitu dugu, jatorrizko sistema gainbegiratu batetik hasita hobetzeko gai dena. Esperimentu simulatuak egin ditugu DoQAren gainean, non erabiltzaileen *feedback*a anotazio estatikoetatik eratortzen dugun. Gure emaitzek erakusten dute gure metodoa gai dela jatorrizko sistema gainbegiratu bat hobetzeko, guztiz gainbegiratua izan den sistema baten errendimendura gerturatuz.

Nahiz eta hasierako emaitza hauek positiboak izan, algoritmoak mugak ditu

gizakien *feedback* bitar zaratatsua erabiltzerakoan. Hau ebazteko, jatorrizko algoritmoa eguneratzen dugu zarata seinalea modelatu ahal izateko. Kasu zaratatsuan lortutako emaitza negatiboek erakusten dute *feedback* bitar zaratatsua erabiltzeko erronka.

Tesi honetan egindako kontribuzioak komunitateak erabili ditu galdera-erantzun motako elkarrizketa sistemak modu orokorrean hobetzeko. Honetaz gain, *feedback* bitarra eta domeinu egokitzapenean ere erabiliak izan dira.

Gaien aurkibidea

Taulen zerrenda

Irudien zerrenda

1. CHAPTER

Introduction

This thesis belongs to the academic field of natural language processing (NLP). Natural language processing is a branch of computer science and artificial intelligence that deals with the interaction between computers and human languages. NLP involves a wide range of tasks, including speech recognition, natural language understanding (NLU) and generation (NLG). The Ixa group inside the HiTZ center in the University of the Basque Country is one of the reference research teams working in NLP. Since the very beginning of the Ixa group more than 30 years ago, it has been a pioneer on the development of NLP tools in many different tasks, paying especial attention to the development of language tools for the Basque language. Moreover, the group participates in European and worldwide level research projects making great contributions for more languages apart from Basque.

This thesis focuses on the interaction between humans and computers in the field of natural language processing (NLP). Specifically, it covers conversational systems and lifelong adaptation, which we analyze by incorporating human feedback into the system over time. The Ixa group has significant experience in these subtasks, as evidenced by their previous work with external collaborators. In a recent study (Agirre *et al.* 2019), the authors define lifelong learning as the ability of dialogue systems to autonomously improve over time. Apart from that, due to the importance of evaluation methods, Deriu *et al.* (2021) presented a survey with various methods for evaluating dialogue systems. The team has also made efforts to apply question answering techniques to both general (Pradel *et al.* 2020) and specific domains (Aceta *et al.* 2021; Otegi *et al.* 2022).

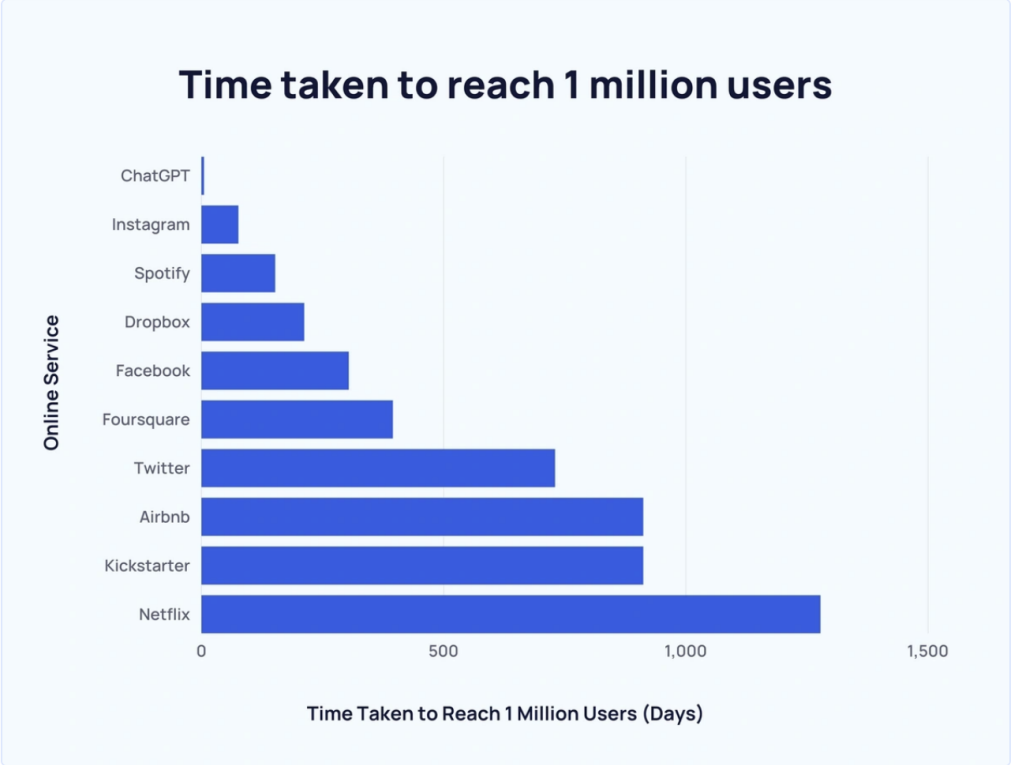
In terms of projects, this thesis work was closely linked to the international project called "Learning to Interact with Humans by Lifelong Interaction with Humans" (LIHLITH). During the course of the thesis, some of the developed research ideas were applied to the "Automated Surveillance of Key Questions on COVID-19 in Scientific Publications" (Vigicovid) project. Additionally, the work done in this thesis, along with the proposed future ideas, was recognized with a Google Faculty Research Award in 2019.

1.1 Motivation

Interpersonal communication is defined as the exchange of information between two or more people (Roloff 1981) and language is a fundamental aspect of it. In recent times, language has also emerged as a vital tool for interacting with machines, as we seek to create more intuitive and natural interfaces that allow us to communicate with computers as easily as we would with another person. One of the key drivers of this trend is the rise of natural language processing (NLP) technologies, which have made significant progress in recent years. Thanks to these advancements, computers can now understand and generate language in ways that were once thought to be impossible (Bubeck *et al.* 2023; Thoppilan *et al.* 2022).

The emergence of conversational systems presents a unique opportunity for human-machine interaction that surpasses traditional methods, such as keyboards. This innovative technology can be applied in various contexts, including hands-free and eye-free interaction, which could be particularly useful in car driving or home control applications. Advancements in speech recognition have led to the popularity of virtual assistants like Amazon Alexa, Google Home, Siri, and Cortana. In the last year, due to the advances in computation (Shoeybi *et al.* 2019; Rasley *et al.* 2020) and data scalability (Gao *et al.* 2020), large language models (LLMs) have become the new paradigm for developing general purpose assistants (Askell *et al.* 2021) that go beyond the previous task-oriented technology (Peng *et al.* 2020). This sudden and unprecedented trend has been demonstrated by the rapid growth of systems such as ChatGPT [1], that currently has over 100 million active users and reached a million users in just 5 days. As shown in Figure 1.1, the growth of ChatGPT outpaces that of major services such as Netflix, Instagram, and Facebook, highlighting the enormous potential of this technology.

[1] `https://chat.openai.com/`



1.1 Figure – Graph showing the time taken by different services to reach 1 million users measured in days. Source:`https://explodingtopics.com/blog/chatgpt-users`.

However, despite the numerous benefits of such systems, there are still several limitations that need to be addressed. One of the main challenges is that these systems generate text that contains missinformation (Lin *et al.* 2021), offensive language (Gehman *et al.* 2020) and factually incorrect statements (Stiennon *et al.* 2020). In recent years, the use of human feedback has been proposed as a solution to improve the performance of NLP systems (Stiennon *et al.* 2020; Christiano *et al.* 2017; Ouyang *et al.* 2022). By allowing users to provide feedback on the correctness and relevance of responses generated by these systems, developers can train the models to better understand and interpret human preferences, leading to better systems. This approach has been shown to be effective in improving the performance of conversational systems and enhancing the overall user experience (Ouyang *et al.* 2022). However, the challenge remains in how to effectively and

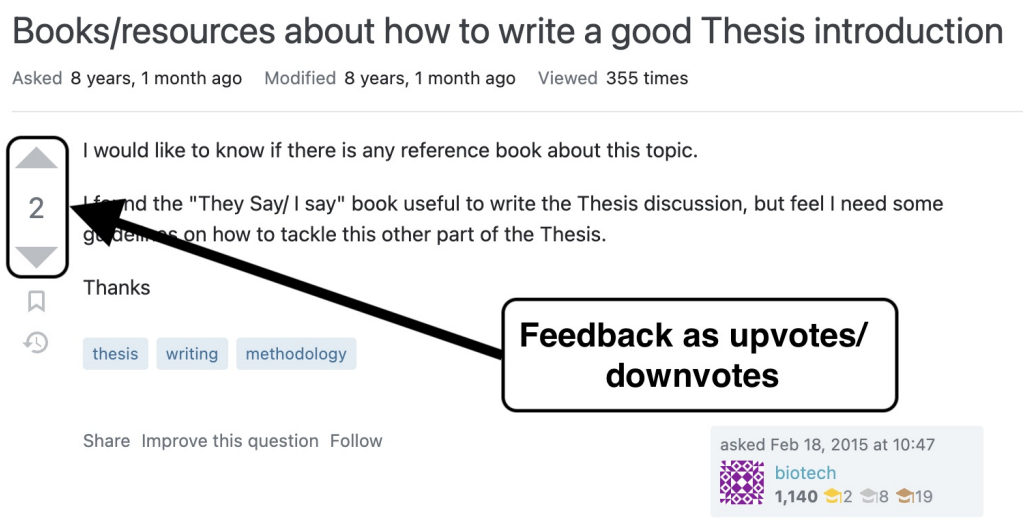
efficiently collect and incorporate feedback from a large and diverse user base, specially when this feedback is noisy and comes from users instead of trained annotators.

To address these challenges, this thesis explores the use of binary human feedback as a means of improving the performance of conversational question answering (CQA) systems. We focus on CQA because it is a well defined task in academia with a well defined evaluation procedure. Moreover, CQA is interactive by nature, so this enables us to focus on the user noisy feedback setting that is one of the main interests of this thesis. Our research focuses on binary feedback as it is the most widely used form of feedback. Many examples on the web require humans to give binary feedback, such as community question answering webpages like *StackExchange*, where users can upvote or downvote the answers written by other users (see Figure 1.2 for an example). Therefore, we assert that binary feedback is the fastest and easiest way for humans to provide feedback. By collecting binary feedback, we aim to train these systems to better understand and respond to natural language queries and to better adapt themselves to new situations, such as out-of-domain conversation. Through a combination of machine learning and natural language processing techniques, we hope to develop more accurate and effective CQA systems that are capable of providing high-quality answers to a wide range of questions in real-world settings. Ultimately, this research could pave the way for a new era of human-machine interaction, where feedback plays an even more significant role in how machines adapt to unseen tasks and user needs.

1.2 Goals and research lines

The main goal of this thesis is to exploit the interaction that deployed systems have with humans by leveraging human feedback as a learning and adaptation signal. In order to achieve this goal we have followed this procedure: (i) create a conversational dataset that reflects the challenges that deployed systems face, (ii) define a lab setting for human feedback simulation and (iii) design a learning algorithm that is able to leverage human simulated feedback in the developed dataset that mimics the difficulties that CQA systems face in reality. More specifically, the research lines developed in this thesis are the following ones:

- **RL1: Creation of a CQA dataset that reflects the challenges faced by deployed systems.** In this research line we have analysed the limitations of state-of-the-art conversational question answering datasets and have pro-



1.2 Figure – Example of a StackExchange post where the upvotes/downvotes, that are a way of binary feedback, can be seen in the top left.

posed a new dataset that addresses those limitations. When we started developing this thesis there were already some public datasets released but they did not capture many of the challenges that real systems face, such as domain shifts.

- **RL2: Development of a human feedback simulation algorithm.** Gathering real human feedback is very expensive, especially in an academic setting where usually no real systems are deployed. In this research line, we presented a user binary feedback simulation algorithm just using a static supervised dataset. We paid special attention to the noise coming from the user by performing a user study task to estimate the amount of noise that a deployed system could expect.

- **RL3: Design of a learning algorithm that leverages binary feedback.** In this research line, we worked on designing an algorithm that was able to adapt a classifier just using simulated binary feedback. An effective algorithm that just uses binary feedback could open new ways to adapt real deployed systems to new tasks, so in this research line we work on the design of such an algorithm and iterate over it to adapt it to new challenges, such as the noisy signals coming from the users.

1.3 Structure of the thesis

This thesis contains two main blocks divided in two languages: Basque and English. The English block is divided along the Chapters 1, 2 and 7. In Chapter 1 we start by presenting the topic of the thesis and the motivation for it (Section 1.1). After that, we explain the goals of the thesis and the different research lines that are part of it (Section 1.2). Later, we enumerate the scientific publications developed during the thesis years (Section 1.4). In Chapter 2 we explain the basics and related work on question answering (Section 2.1) and learning from feedback (Section 2.2). In Chapter 7, that is the last English chapter, we present the main conclusions, contributions and future work drawn from this thesis work.

In the Basque block we present the main chapters of the thesis, where the developed projects are explained. This block is divided in Chapters 3, 4, 5 and 6. In Chapter 3 we present the conversational question answering dataset that we have created, where we explain the methodology we followed and the properties of it. Later, in Chapter 4 we show the main experiments and results on this dataset, especially focusing on domain transfer and usage of conversational context. In the last Chapters 5 and 6 we analyze the learning from binary user feedback settings, where we present two scenarios, a simple one without user noise (Chapter 5) and a more realistic one that tries to capture the noise coming from the users (Chapter 6).

1.4 List of scientific contributions

In this section we present the scientific contributions that have been developed during the years of the thesis. This section is split in two parts: first we present the publications that are part of the manuscript and then we show the ones that are not part of it. All the papers are presented in chronological order.

Contributions that are part of the thesis

Campos et al. (ACL 2020) presented in Chapters 3 and 4

Campos, J. A., Otegi, A., Soroa, A., Deriu, J. M., Cieliebak, M., & Agirre, E. (2020). DoQA: accessing domain-specific FAQs via conversational QA. In 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), online, 5-10 July 2020 (pp. 7302-7314). Association for Computational Linguistics. Citations: 40.

Campos et al. (COLING 2020) presented in Chapter 5

Campos, J. A., Cho, K., Otegi, A., Soroa, A., Agirre, E., & Azkune, G. (2020, December). Improving Conversational Question Answering Systems after Deployment using Feedback-Weighted Learning. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 2561-2571). **Outstanding paper mention.** Citations: 6.

Contributions that are not part the thesis

Agerri et al. (LREC 2020)

Agerri, R., San Vicente, I., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., & Agirre, E. (2020, May). Give your Text Representation Models some Love: the Case for Basque. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 4781-4788).

Otegi et al. (LREC 2020)

Otegi, A., Agirre, A., Campos, J. A., Soroa, A., & Agirre, E. (2020, May). Conversational question answering in low resource scenarios: A dataset and case study for basque. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 436-442).

Deriu et al. (EMNLP 2020)

Deriu, J. M., Tuggener, D., von Däniken, P., Campos, J. A., Rodrigo, Á., Belkacem, T., & Cieliebak, M. (2020, November). Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 3971-3984). **Nomination to best paper award.**

Otegi et al. (EMNLP Workshop 2020)

Otegi, A., Campos, J. A., Azkune, G., Soroa, A., & Agirre, E. (2020, December). Automatic evaluation vs. user preference in neural textual QuestionAnswering over COVID-19 scientific literature. In Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020.

Salaberria et al. (Ikergazte 2021)

Salaberria, A., Campos, J. A., Garcia-Ferrero, I., & Fernandez de Landa, J. (2021, June). Itzulpen Automatikoko Sistemen Analisia: Genero Alborapenaren Kasua. In Proceedings of the IV. Ikergazte. Nazioarteko ikerketa euskaraz. Kongresuko artikulu bilduma. Ingeniaritza eta Arkitektura.

Fernandez de Landa et al. (Ikergazte 2021)

Fernandez de Landa, J., Garcia-Ferraro, I., Salaberria A., & Campos, J. A. (2021, June). Twitterreko Euskal Komunitatearen Eduki Azterketa Pandemia Garaian. In Proceedings of the IV. Ikergazte. Nazioarteko ikerketa euskaraz. Kongresuko artikulu bilduma. Ingeniaritza eta Arkitektura.

Scheurer et al. (ACL Workshop 2022)

Scheurer, J., Campos, J. A., Chan, J. S., Chen, A., Cho, K., & Perez, E. (2022). Training language models with language feedback. In The First Workshop on Learning with Natural Language Supervision at ACL.

Chen et al. (Under review for ICML 2023)

Chen, A., Scheurer, J., Korbak, T., Campos, J. A., Chan, J. S., Bowman, S.R., Cho, K., & Perez, E. Improving Code Generation by Training with Natural Language Feedback.

Scheurer et al. (Under review for ICML 2023)

Scheurer, J., Campos, J. A., Korbak, T., Chan, J. S., Chen, A., Cho, K., & Perez, E. Training Language Models with Language Feedback at Scale

Garcia-Ferrero et al. (Accepted at SemEval 2023)

Garcia-Ferraro, I., Campos, J. A., Sainz, O., Salaberria, A., & Roth, D. Context-enriched Multilingual Named Entity Recognition using Knowledge Bases

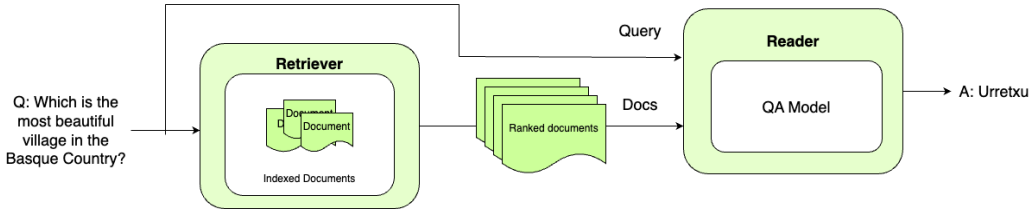
2. CHAPTER

Background

The main aim of this thesis is to develop interactive systems that adapt and continue improving after deployment. In order to achieve it, the thesis builds on question answering (QA) systems and feedback. The motivation for that is that QA systems and specially the conversational ones are interactive by default. Also, feedback is the most natural way of supervision once the systems have been deployed. Due to that, in this chapter we introduce the basis of these two topics, defining the most important notions in order to follow the thesis. It is important to mention that QA and learning from feedback have a very long tradition in NLP and that they are still main research topics nowadays. So, the intention of this chapter is not to make fine-grained analysis of these topics and to present every single technique applied, instead we will just present the techniques that we use later during the thesis.

2.1 Question answering systems

There are a wide variety of ways for classifying QA systems (Mishra and Jain 2016). According to Jurafsky and Martin (2014) traditionally there have been two major paradigms for QA: knowledge based and information retrieval based models. With the emergence of large generative language models (LLMs) and their impressive performance on a wide range of NLP tasks, non-retrieval question answering is becoming more and more popular (Brown *et al.* 2020; Zhang *et al.* 2022; Scao *et al.* 2022; Black *et al.* 2022). This type of systems suffer from



2.1 Figure – The general architecture of an ODQA system. The retriever receives a query and ranks the collection of documents. The reader input consists of a query and the ranked documents and returns an answer.

hallucination (Ortega *et al.* 2021), but LLMs are still able to answer questions without the retrieval step due to their capability to encode knowledge. In this thesis, we will include information retrieval based models inside the broader open-domain QA (ODQA) models. We make this modification as ODQA is a more general term that includes retrieval and non retrieval based methods.

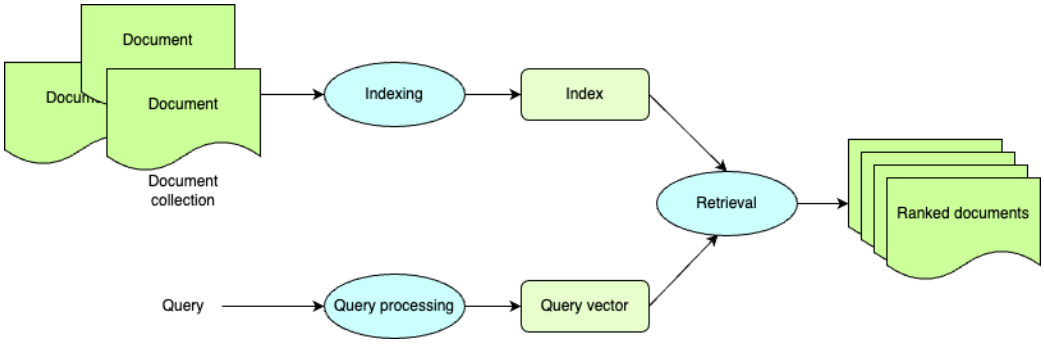
Knowledge-based QA involves mapping a natural language query to a structured knowledge base using a query language such as SQL for relational databases or SPARQL for simpler databases like RDF triplets:

"Who did write the *Gernikako Arbola* song?" $\rightarrow$ artist (Gernikako Arbola, ?x)

However, as knowledge-based QA systems are not a main research topic of this thesis, we just focus on the ODQA paradigm.

2.1.1 Open-domain question answering

The previously mentioned knowledge-based systems use structured data for answering questions. Using structured data for this task allows to build quite reliable systems. But, as the amount of information stored in structured ways is limited, the idea of open-domain QA comes up. The most common approach for ODQA is Information-retrieval or IR-based QA. IR-based QA has two main steps. In the first one, given a user question, IR techniques are used to collect relevant documents and passages from the web. The general architecture of ODQA can be seen in Figure 2.1. The second step of IR based QA consists of using machine comprehension algorithms for understanding the content of the relevant documents and trying to answer the user questions afterwards.



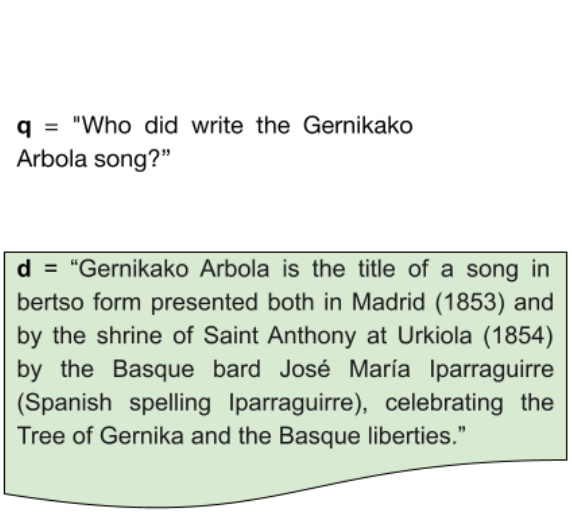
2.2 Figure – The general architecture of an information retrieval system. The queries and documents are processed for generating the query vector and document index. In the retrieval step, given the processed index and query vector the documents are ranked.

2.1.2 Information-retrieval

Information retrieval encompasses the retrieval of any kind of media based on user information needs. Among the different media, we focus on text and specially in the case in which a user poses a query to a retrieval system and the system returns a set of documents from a collection. We consider any unit of text as a *document*, such as a web page, a scientific paper or a very short passage from Wikipedia. The *collection* is the set of documents used to satisfy the user requests and this can be as big as the whole web for the case of search engines. The high level architecture of an IR system can be seen in Figure 2.2. Here, starting from the collection of documents and the user query we first index all the documents and the query by obtaining a vector representation of each of them. Later, we perform the retrieval step in which using a distance metric we return a list of ranked documents taking the relevancy of them into account for the specific query.

Vector representations for IR

Vector representations can be mainly divided into two categories: sparse representations and dense representations. The sparse representations are the most traditional ones and have been broadly used for many years (Robertson *et al.* 2009; Géry and Largeron 2012). The most simple sparse vector space model for IR is a bag-of-words model where queries q_i and documents d_i are mapped to vectors using unigram word counts (Salton 1972). An example of this method can be seen in Figure 2.3.



Words	q	d
the	1	5
of	0	3
(	0	3
)	0	3
in	0	2
and	0	2
by	0	2
Basque	0	2
Iparraguirre	0	2
Gernikako	1	1
Arbola	1	1
…	…	…

2.3 Figure – Example of unigram word counts for query q and document d representations. The queries and documents are first tokenized and then the unigrams are counted for generating sparse vectors.

When generating sparse vector representations for IR, term weights are used instead of raw word counts. The most common term weight schema is known as tf-idf. Tf-idf is the product of the term frequency tf and the inverse document frequency idf. Term-frequency is usually computed using the logarithm of the count of each term t in the document d: $tf_{t,d} = log(\text{count}(t, d) + 1)$. On the other hand, the document frequency of a term is the count of the documents in which the term appears. The uniqueness of the terms make them useful for discriminating documents, which is specially useful for IR. Sparck Jones (1972) defined the inverse document frequency as presented in Equation 2.1, where N is the total amount of documents in the collection.

$$idf_t = \log_{10} \frac{N}{df_t} \tag{2.1}$$

However, this sparse representations have a well known conceptual flaw: they only work when there is an exact overlap of words between the query and the document. This issue is known as the vocabulary mismatch problem (Furnas *et al.* 1987). Dense vector representations have been proposed as the solution for the vocabulary mismatch problem. In (Deerwester *et al.* 1990) the authors presented LSI as a first approach, in which they used singular value decomposition (SVD) for approximating the term-weight matrix with one of a lower rank. After the success of pre-trained language models (PLM)s, more modern techniques use encoders like BERT (Devlin *et al.* 2019) for getting the vector representations. There has been an explosion of methods that use PLMs for IR (Zhao *et al.* 2022). Two main architecture types have been used for PLM based dense representations: single representations, where the query and the document are encoded using the same LM (Guu *et al.*, 2020; Izacard *et al.*, 2021; Xu *et al.*, 2022a; Dai *et al.*, 2023; Gao and Callan, 2022) and dual representations (Humeau *et al.* 2020; Karpukhin *et al.* 2020; Khattab and Zaharia 2020), where the query and the document are encoded using different LMs. Even if dense vector representations have shown strong performance, more traditional sparse representations are still very strong baselines (Thakur *et al.* 2021).

Document scoring

Given a collection of documents $D = \{d_1, d_2, \cdots, d_n\}$ and a query q, the most common way of scoring them is to use the *cosine similarity* on top of their vector representations:

$$score(q, d_i) = cos(q, d_i) = \frac{q \cdot d_i}{|q||d_i|}$$

Instead of *cosine similarity*, simple *dot product* is also used as a scoring function. Based on tf-idf, Robertson *et al.* (1995) presented the BM-25 ranking function. This new algorithm takes into account the length of the document and the average document length in the collection. The BM25 score for a given query term q and document d_i is given by:

$$score(q, d_i) = \sum_{t \in q_i} \underbrace{\log(\frac{N}{d_i f_t})}_{\text{IDF}} \underbrace{\frac{tf_{t,d_i}}{k(1 - b + b(\frac{|d_i|}{|D_{avg}|})) + tf_{t,d_i}}}_{\text{weighted tf}} \tag{2.2}$$

where tf_{t,d_i} is the frequency of term t in document d_i, $|D_{avg}|$ is the average document length in the collection, $|d_i|$ is the length of document d_i and k and b are free parameters that control the impact of term frequency and document length normalization. The BM25 score is computed for each query term and then aggregated to obtain the final ranking score for the document. Higher scores indicate that a document is more relevant to the query. The choice of the parameters k and b can have a significant impact on the ranking performance of the BM25 algorithm and their optimal values may depend on the characteristics of the collection and the query set.

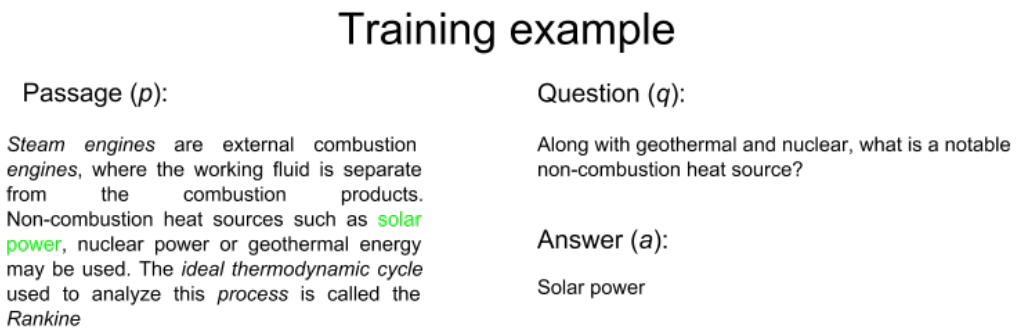
Evaluation

Similar to many other machine learning (ML) tasks, precision and recall are the main metrics used in information retrieval. In order to compute these metrics, we define every document in our collection as relevant or not relevant. After making this assumption, precision will represent the fraction of the returned documents that are relevant and recall will represent the fraction of all the relevant document that are returned.

Precision and recall are not the ideal metrics for benchmarking IR rankings since they do not give higher scores to the method that ranks relevant documents higher. In order to approach this issue mean average precision (MAP) is proposed as a solution. Given R_r the set of relevant documents and $Precision_r(d)$ the precision in the rank at which relevant document d is found, the average precision (AP) is calculated in the following way:

$$AP = \frac{1}{|R_r|} \sum_{d \in R_r} Precision_r(d) \tag{2.3}$$

Then, MAP is just the average of the AP of all the queries.



2.4 Figure – A training example taken from the SQuAD dataset (Rajpurkar *et al.*, 2016). This example contains a passage p, a question q and an answer a, which is a subspan of the passage.

2.1.3 Machine reading comprehension

Teaching computers how to read and understand the meaning of text has always been one of the major research goals of NLP. In order to achieve this goal, NLP researchers have performed machine reading comprehension (MRC) research for many years. As early as 1977, Lehnert (1977) built a QA system named QUALM that was able to read stories and answer questions about what it read. After this initial system, many rule based and statistical methods followed (Hirschman *et al.* 1999; Charniak *et al.* 2000; Riloff and Thelen 2000). At this time, the bag of words (BoW) was the most common representation for the text. Given a question q and a target passage with sentences $s_1, ...s_n$ the sentence that has the greatest intersection with the question was selected as the answer. Charniak *et al.* (2000) replaced the BoW schema with the tf-idf statistic and showed significant improvement. However, as in any machine learning task, it is hard to make great progress in the field without good datasets and there was a big lack of datasets at this time. Due to this, the research on MRC slowed down until 2016 when Rajpurkar *et al.* (2016) presented the first large scale dataset. See Figure 2.4 for an example of the dataset.

MRC is formulated as a supervised learning problem. Given a dataset containing triplets $\{(p_i, q_i, a_i)\}$ the goal is to train a predictor f that given a passage p_i and a question q_i has to output the answer a_i.

$$a_i = f(p_i, q_i)$$

Conversational QA is a special case of MRC that follows up the idea by modifying the problem formulation to support a dialogue between the questioner and the system. In this setting, the system is required to answer a sequence of questions in the context of a conversation by attending to a passage. Specifically, given a passage p_i, a conversation history $H_i = (q_{i,1}, a_{i,1}), (q_{i,2}, a_{i,2}), ..., (q_{i,k-1}, a_{i,k-1})$ and a new question $q_{i,k}$, the goal is to predict the answer $a_{i,k}$.

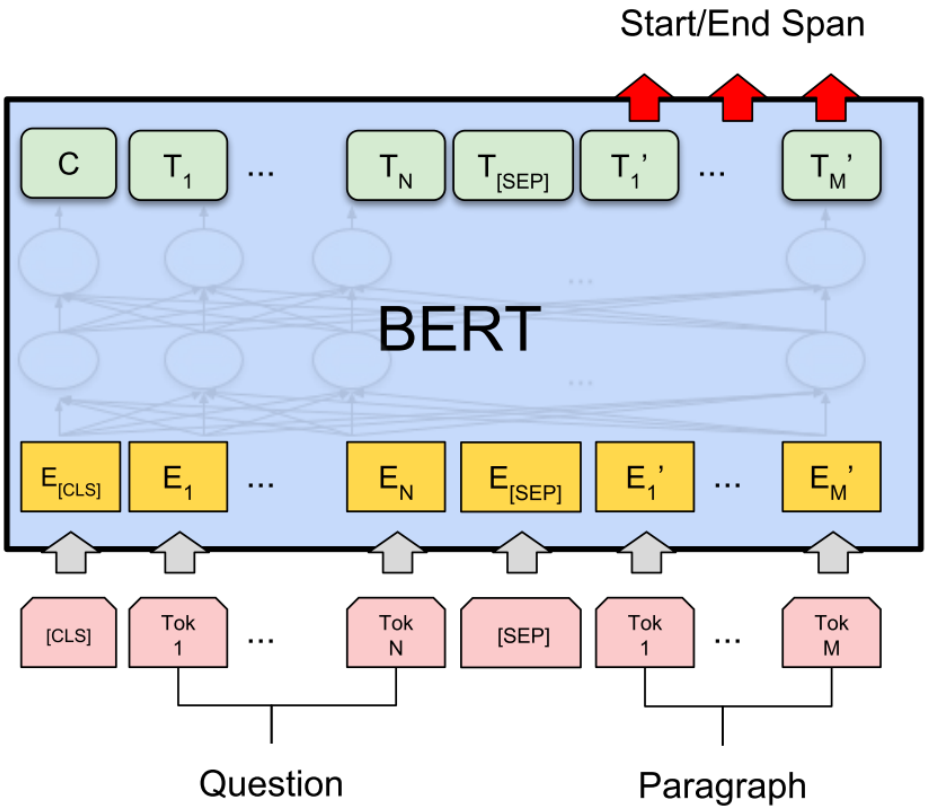
$$a_{i,k} = f(H_i, p_i, q_{i,k})$$

Usually, MRC tasks are divided into four categories: close style, multiple-choice, span prediction and free-form answer (Chen 2018; Qiu *et al.* 2019; Liu *et al.* 2019a). Among all those, in this thesis we will specially focus on the span prediction tasks as these are the ones we used during the development of the thesis project.

Span prediction

The objective of span prediction is to find a span in the passage that answers the question posed by the user. More specifically, given a document or passage p consisting of M tokens $p_1, p_2, ..., p_m$ and a user question q consisting of N tokens $q_1, q_2, ..., q_n$. Our task is to locate an answer, denoted as a, that is part of p. To achieve this, we compute the start $p_{start}(i)$ and end $p_{end}(i)$ probabilities for each token i in p. The value of $p_{start}(i)$ indicates the probability of token i being the starting token of the answer, while $p_{end}(i)$ indicates the probability of token i being the ending token of the answer.

Similar to other NLP tasks, feature-based methods have been substituted by neural network approaches in the span prediction task over the past few years. These approaches typically use deep learning architectures, such as recurrent neural networks (RNNs) (Wang and Jiang 2017; Seo *et al.* 2017; Huang *et al.* 2018; Yatskar 2019; Peters *et al.* 2018) and more recent transformer-based models like BERT and its variants that follow the masked language modelling (MLM) pre-training (Devlin *et al.* 2019; Liu *et al.* 2019b; Clark *et al.* 2020; Lan *et al.* 2020). Pre-trained language models (PLMs) as BERT are trained on large amounts of text data and then fine-tuned on the specific task of MRC to learn representations that capture semantic and syntactic information in the input text. Pointer networks (Vinyals *et al.* 2015) are the most common approach for fine-tuning PLMs on the MRC task. More specifically, two new vectors $S, E \in \mathbb{R}^H$ are included at fine-tuning time for calculating the following probabilities:



2.5 Figure – Example of the BERT architecture for QA. Here the question q and the paragraph or passage p are fed into the system using the [SEP] token as separation. For the output, the start and end probabilities are computed on top of the passage tokens. Source: (Devlin *et al.* 2019)

$$p_{start}(i) = \frac{e^S \cdot T_i}{\sum_j e^S \cdot T_j} \; p_{end}(i) = \frac{e^E \cdot T_i}{\sum_j e^E \cdot T_j}$$

where T_i is the last hidden representation of the p_i token in the LM. See Figure 2.5 for a visual example. On top of this architecture, recent studies have explored various techniques to improve the robustness and efficiency of these models, such as incorporating external knowledge (Qiu *et al.* 2019; Yang *et al.* 2019; Pan *et al.* 2019) and using data augmentation techniques (Puri *et al.* 2020; Shakeri *et al.* 2020).

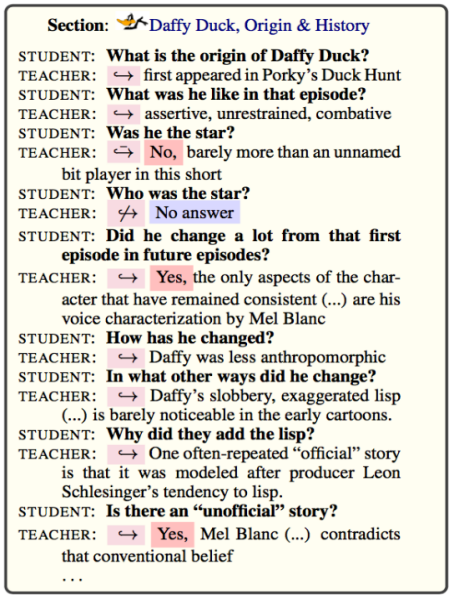
In the case of CQA, different efforts have been performed for modelling the context (Qu *et al.* 2019; Chen *et al.* 2021; Huang *et al.* 2019) and enriching the questions with the context (Lin *et al.* 2020; Vakulenko *et al.* 2021; Elgohary *et al.* 2019) on top of BERT like PLMs. Due to the success of generative LMs (Sanh *et al.* 2022; Lewis *et al.* 2020; Raffel *et al.* 2020) they have been proposed as an alternative to pointer networks showing very strong performance (Wang *et al.* 2019b: a). However, as the gain on popularity of these generative PLMs came after the main developments of this thesis we do not further analyze them.

Datasets

Recent work on the field has resulted in the creation of multiple large scale datasets (Rajpurkar *et al.* 2016; Trischler *et al.* 2017; Nguyen *et al.* 2016; Kočiský *et al.* 2018; Kwiatkowski *et al.* 2019; Dunn *et al.* 2017; Yang *et al.* 2018; Castelli *et al.* 2019; Nguyen *et al.* 2016). Typically, these datasets consist of multiple pairs of questions and answers, with a reference passage provided for each answer. While the questions are always presented in free text form, most of the answers are given as a contiguous span from the reference passage and just a few cases provide free-form answers. Extractive datasets refer to the former, while abstractive ones refer to the latter. Overall, these QA datasets do not involve any dialogue structure as the queries are unrelated to each other.

A key consideration when using these datasets is to know their differences. One important distinction is between large-scale datasets in general domains, such as SQuAD (Rajpurkar *et al.* 2016) and SQuAD2.0 (Rajpurkar *et al.* 2018), and domain-specific datasets, such as TechQA (Castelli *et al.* 2019), which focus on technical support. Another important difference is whether the questions and answers were generated by crowdworkers with access to the passage (e.g., SQuAD, SQuAD2.0), or if they come from real user queries in search engines (Nguyen *et al.* 2016; Kwiatkowski *et al.* 2019). The latter ensures that the questions are representative of real-world queries. To increase the difficulty of MRC, there have been efforts to develop datasets requiring multiple documents to answer questions, such as HotpotQA (Yang *et al.* 2018).

This thesis is primarily focused on conversational QA, and as such, it is important to explore the existing datasets developed for this subtask in MRC. Two prominent conversational QA datasets are CoQA (Reddy *et al.* 2019) and QuAC (Choi *et al.* 2018), which consist of QA dialogues designed to satisfy the information needs of a user by answering questions on a range of topics. Both datasets are created through crowdsourcing, where a questioner is given a topic and asked



2.6 Figure – Conversation example from the QuAC dataset. In this example, a conversation about the Daffy Duck can be seen. The arrows in the teacher turns are used to visualize the continuation dialog acts. These dialog acts are used to guide the student through the conversation. Source: (Choi *et al.* 2018)

to pose open-ended questions about it. An answerer then selects an answer to the question by choosing an excerpt from the relevant passage that describes the topic. Some questions in both datasets are intentionally unanswerable, which increases the difficulty of the task. Moreover, due to the conversational nature of the datasets, access to the conversational history is required in order to answer most of the questions. Figure 2.6 shows an example of a conversational from the QuAC dataset.

CoQA consists of 127k questions with answers derived from 8k conversations on passages covering various domains from children stories to science. The answers provided by the answerer are excerpts from the relevant passage, which can be reformulated as per their preference. The authors report that 78% of the answers had at least one edit. Despite the benefits of reformulating answers for natural dialogues, Yatskar (2019) found that span-based systems can achieve a performance of up to 97.8 F1 score, indicating that editing answers does not lead to systems with better quality. In CoQA, both the questioner and answerer have access to the full passage, which guides the conversation towards the specific in-

formation conveyed in it.

QuAC is a dataset that consists of 14k information-seeking QA dialogues. These dialogues are centered around a specific section in Wikipedia articles related to people. The answerer in QuAC has access to the full section text, while the questioner can only view the section's title and the first paragraph of the article as inspiration for formulating queries. In addition to the question-answer pairs, QuAC includes dialogue acts in each turn, enabling the answerer to inform the questioner whether to continue asking questions related to the previous answer or to shift the focus to other aspects of the topic.

After the publication of the dataset developed during this thesis, there has been a surge in CQA dataset creation efforts, with a particular focus on ODQA tasks (Feng *et al.* 2020; Qu *et al.* 2020; Feng *et al.* 2021; Anantha *et al.* 2021; Guo *et al.* 2021; Adlakha *et al.* 2022). Among these datasets, TOPIOCQA (Adlakha *et al.* 2022) is the most complete, as it contains topic-switching dialogues while maintaining the interesting properties of previous datasets, namely its multi-turn, open-domain, contains free-form answers, and information-seeking questions. Despite having only 3,920 dialogues, which is smaller than CoQA and QuAC, TOPIOCQA's dialogues are longer due to its ability to accommodate topic changes within the same conversation.

It is essential to emphasize the critical role that good benchmarks play in developing and evaluating conversational QA systems. They provide a common ground for researchers to compare and analyze different approaches. Therefore, generating challenging and realistic datasets has become increasingly important to advance any task in this field. This explains the significant efforts put into creating the most comprehensive datasets possible.

Evaluation

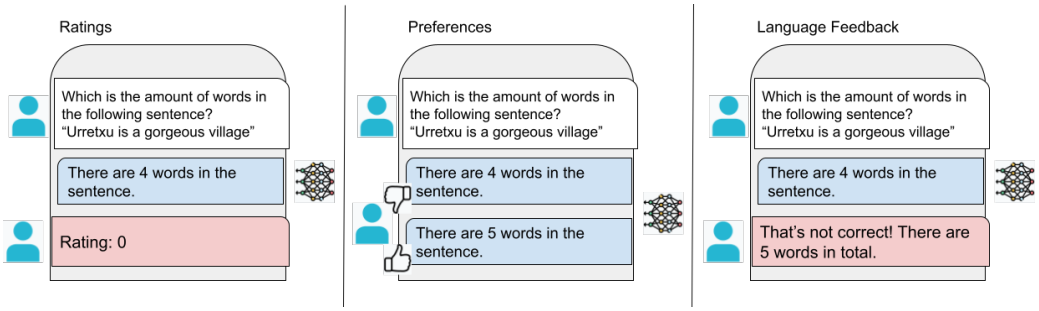
The most expanded metric for evaluating MRC systems is the token level F1 metric. In order to calculate the F1, the overlap between the system answer tokens $a_i = \{a_{i1}, a_{i2}, ..., a_{ik}\}$ and the annotated answer tokens in the dataset $\hat{a}_i = \{\hat{a}_{i1}, \hat{a}_{i2}, ..., \hat{a}_{il}\}$ is calculated. In the process of calculating the overlap, answers are treated as bag of words and the average F1 of all the answers in the dataset is reported as the final metric. Since many MRC datasets contain multiple correct answers, the maximum F1 per answer is always considered. Exact match (EM) is also very common on QA. EM measures the percentage of questions where a_i perfectly matches $\hat{a}_i$. In the case of CQA, the human equivalence score on a question level (HEQ-Q) is usually used for evaluation too. This metric

calculates the proportion of questions in which the F1 obtained by the system is better than the F1 obtained by humans. The human F1 score is measured thanks to the multiple answers that the MRC datasets contain.

2.2 Learning from feedback in NLP

In the field of machine learning, defining the desired objective for learning algorithms can often be challenging. For example, when training models to provide accurate answers to questions, we use a proxy objective of training models to answer questions similar to how humans would, by providing demonstrations. Although demonstrations have been the most common form of human supervision, they are static and do not account for the different domains that the system may encounter at deployment. This process, also known as supervised learning, has two main problems. Firstly, there is a misalignment between the proxy objective of answering questions like humans do in demonstrations and the real objective of providing correct answers. This can make it challenging for NLP systems to provide reliable answers to questions that humans struggle to answer. Secondly, it is challenging to capture all the different domains that the system will encounter at deployment time in a static dataset. As a result, the NLP system may struggle to be reliable in those out-of-domain cases.

As we previously mentioned, pre-trained language models have become the new de facto standard for solving NLP tasks. PLMs are usually trained using learning objectives that leverage unlabeled text. For example, PLMs are often trained to predict the next word given some large corpora that is usually mined from the internet (Radford *et al.* 2018: 2019). Masked language modeling (MLM) is another common approach for training PLMs that consists of predicting words that have been masked out (Devlin *et al.* 2019; Liu *et al.* 2019b). These two learning objectives are powerful because they give PLMs the capacity of learning from vast amounts of text. Language models that have been trained following this objectives achieve strong performance across many NLP tasks, ranging from summarization to QA (Brown *et al.* 2020; Rae *et al.* 2021; Stiennon *et al.* 2020). However, the learning objectives for PLMs training is also a proxy objective; the training text is human-written and it even if PLMs should learn to imitate it, we can find many undesired behaviours in aggressively mined internet data (Gao *et al.* 2020). Some examples of problematic PLMs outputs are generating miss information (Lin *et al.* 2020), offensive language (Xu *et al.* 2021; Gehman *et al.* 2020) or factually incorrect text (Stiennon *et al.* 2020).

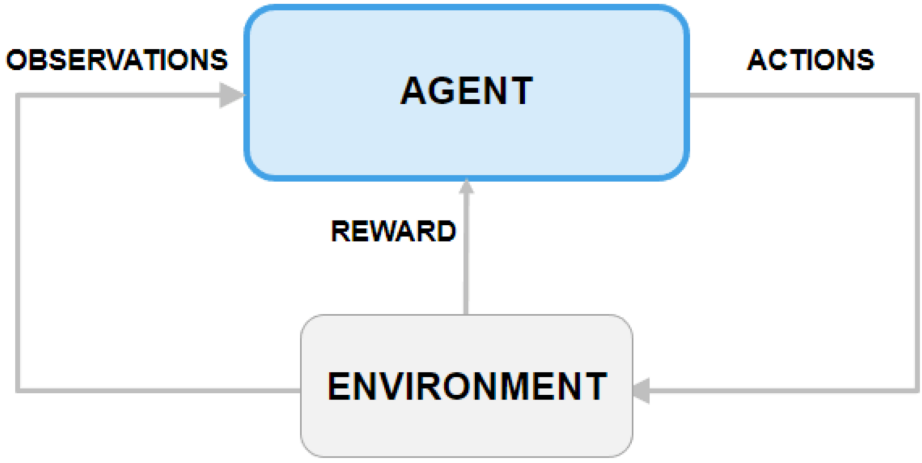


2.7 Figure – Most common ways for giving feedback to ML outputs. In the case of ratings, humans give a scalar score to each of the ML outputs. In the preferences, given two ML outputs the human select the preferred one. For the language feedback signal, an explanation of why the output is not correct is given.

In order to address the problems mentioned earlier, one potential solution is to use human feedback. Human feedback comes in different forms, but the most common types are (see Figure 2.7 for a visual example):

- Ratings: given y_i the output generated by a PLM, the annotator gives a scalar score based on how accurate it is. Binary feedback is a special case of this setting where the score is either 1 or 0.
- Preferences: given two outputs $\{y_{i1}, y_{i2}\}$ generated by a PLM, the annotator has to select which of the outputs is more correct. This is similar to binary feedback, but requires the annotator to compare multiple outputs.
- Language feedback: given y_i the annotator generates f_i that is a natural language explanation of how the output could be improved. If the output is correct, the feedback might simply be "The answer is correct." There is also the case in which f_i could contain the correct answer directly, in a more similar way to supervised learning.

Many of the methods that leverage feedback as a learning signal are formalized as reinforcement learning (RL) problems. Reinforcement learning is a type of machine learning that involves training an agent to make decisions following a policy π in an environment by maximizing a reward signal. The agent learns by interacting with the environment and receiving feedback in the form of rewards or penalties for its actions. Basic reinforcement learning is modeled as a Markov decision problem (MDP), which is defined as a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is

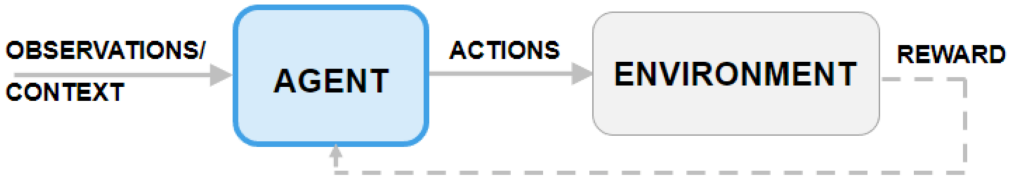


2.8 Figure – Example of the typical framing of a RL scenario. Here an agent takes actions in an environment and it gets a reward. RL is usually seen as a sequencial decision making problem, where each of the actions taken by the agent affects the next observation given by the environment.

the state space, $\mathcal{A}$ is the action space, P is the transition probability function, R is the reward function, and γ is the discount factor. At each time step t, the agent observes the current state s_t, selects an action a_t from the set of available actions $\mathcal{A}$, and receives a reward r_t and a new state s_{t+1} from the environment. The goal of the agent is to learn a policy π that maximizes the expected cumulative reward over a finite or infinite time horizon. A visual example of the general RL setup can be appreciated in Figure 2.8.

2.2.1 Learning from ratings

Learning from ratings feedback is usually formalized as a contextual multi-armed bandit problem. Contextual multi-armed bandits belong to a specific category of reinforcement learning (RL) methods, where at each time step t, the learner is presented with a context or input x_i, selects an action y_i based on a policy π, and receives a reward r_i for the chosen action. In this RL approach, the environment does not provide the agent with a new state. The aim of the learner is to maximize the cumulative reward or minimize the cumulative regret, which is defined as the difference between the learner's and the optimal policy's actions. The regret at time T can be calculated using the following formula:



2.9 Figure – Example of the typical framing of a contextual bandit scenario. Here an agent takes actions in an environment and it gets a reward, but it is not framed as a sequential decision making problem. In other words, the actions taken by the agent do not influence the observations as in traditional RL.

$$R_T = \sum_{t=1}^{T} r_t^* - \sum_{t=1}^{T} r_t$$

where r^* represents the reward the optimal policy would receive. The main challenge in multi-armed bandits is to balance exploration and exploitation while minimizing the overall regret. In conversational systems, x_i includes the conversational history and the current user turn. The conversational model determines the policy π, which is updated to maximize the total reward. An example of the contextual multi-armed scenario can be found in Figure 2.9.

In the work presented by Liu *et al.* (2018), the authors applied multi-armed bandits to dialogue response selection. They used a customized version of the Thompson sampling algorithm (Thompson 1933) to explore contextual multi-armed bandits. The Ubuntu Dialogue Corpus (Lowe *et al.* 2015) was used as a source of technical dialogues for user simulation. In this simulation, retrieving the correct response resulted in positive feedback while any other response resulted in negative feedback. In a similar dialog setting, the authors in (Weston 2016; Li *et al.* 2017) used reward-based imitation (RBI) for training the conversational model to imitate answers that received positive feedback.

After the work presented in this thesis was completed, new efforts followed up. In (Gao *et al.* 2022), the same binary feedback user simulation algorithm proposed in the thesis was used to simulate binary user feedback in the MRC task. They first trained an initial supervised system with limited training data and updated it later leveraging binary feedback simulated on a static supervised dataset. They used a policy gradient method similar to REINFORCE (Sutton *et al.* 1998), but instead of Monte Carlo sampling, they used *arg max* to sample from

the classifier's predictive distribution. They showed that their models were able to handle some amount of noise (up to 20%) depending on the QA dataset they were using and the quality of the initially deployed system.

In the case of non-binary ratings, Li *et al.* (2022) developed a dataset named FEEDBACKQA for retrieval-based Question Answering (QA). This dataset contains user ratings for responses retrieved by an initially deployed QA system. The researchers developed a generative re-ranker utilizing these ratings, which calculates the probability each candidate answer has for receiving an "excellent" rating. The final score for each answer is a linear combination of the originally deployed model and the re-ranker probabilities. Through experiments, they demonstrated the efficacy of training a re-ranker with human ratings to improve information retrieval (IR)-based QA.

2.2.2 Learning from preferences

Reinforcement learning from human feedback (RLHF) is a common method used to address undesired behaviors of PLMs. This method leverages comparisons from users and was first proposed by Christiano *et al.* (2017). After the development of this thesis, the RLHF method has subsequently been applied in many other works (Stiennon *et al.* 2020; Ouyang *et al.* 2022; Bai *et al.* 2022). RLHF can be defined as a three-step process:

1. **Collect samples and annotate them.** For each input x the LM is used to sample $\{y_{i1}, y_{i2}\}$ output pairs. These pairs are then sent to human annotators who provide a preference feedback for one of the two outputs.

2. **Learn a reward model from comparisons.** After generating the human comparisons feedback dataset, a reward model is trained to predict the log odds that the given model output y_i is the preferred one.

3. **Optimize a policy with the reward model.** Finally, a reinforcement learning algorithm is used to optimize the LM by using the output logit of the reward model as a reward.

The main objective of this process is to update the LM in a way that generates higher quality outputs for humans. The most extended reinforcement learning algorithm for achieving this goal is PPO (Schulman *et al.* 2017), where the reward is defined as follows:

$$R(x, y) = r_\theta(x, y) - \beta \log[\pi(y|x)/\pi_0(y|x)]$$

here $r_\theta(x, y)$ is the reward for an entire output y given by the reward model, π_0 is the initial frozen LM, π is the LM being updated and β controls the KL penalty. The KL penalty term between π_0 and π is introduced for encouraging the policy π to explore and avoiding overoptimization.

As we already mentioned, RLHF is a widely used technique for learning from human feedback and has been found to be effective. However, preference evaluation often suffers from ambiguity, as it does not provide reasons for why a particular output is preferred. This ambiguity can lead to issues such as goal misgeneralization (Di Langosco *et al.* 2022). Additionally, Perez *et al.* (2022) has demonstrated cases of inverse scaling in RLHF, where increased use of RLHF leads to a decline in the quality of language models. Examples of this phenomenon include cases in which the LMs showed strong political views or desires to avoid being shut down.

2.2.3 Learning from language feedback

Motivated by the limitations on learning from ratings and preferences, learning from language feedback has been proposed as a potential solution. Language feedback is a more natural and information-rich form of human feedback that conveys more bits of information, enabling a more nuanced and comprehensive understanding of human preferences.

In the context of dialog systems, the authors of (Hancock *et al.* 2019) proposed a chatbot that can detect when it has generated an incorrect response by analyzing user reactions. When the chatbot detects an error, it asks the user for the correct response and adds it to a static dataset that will user for later re-training. This method is effective for improving chatbots, but obtaining the necessary human demonstrations can be difficult in practice. After the work in this PhD was completed, Xu *et al.* (2022b) performed imitation learning on free-form feedback which showed promising results. However, free-form feedback alone is not as a direct training signal as human demonstrations are, so the results are worse in this case. Inspired by the method proposed by Scheurer *et al.* (2022), where initial model outputs are refined using free-form user feedback, Shi *et al.* (2022) modify the previously performed imitation learning approach with a refinement generation algorithm that improves incorrect outputs using free-form feedback. The

resulting refined outputs are then used in imitation learning to further improve the model.

The high cost of collecting human language feedback has led to an increased focus on automated methods for generating language feedback. In a recent study, Saunders *et al.* (2022) demonstrated that language models can produce high-quality feedback for model outputs, which can improve human-generated feedback as well. Another study by Bai *et al.* (2022) directly used LM-generated feedback to enhance a dialogue assistant, eliminating the need for expensive human feedback. In a similar way, Madaan *et al.* (2023) demonstrated that LMs can provide feedback to themselves, without the need for a more powerful LM as the feedback generator. Overall, model generated language feedback is a new area of research and it is still in its infancy.

3. KAPITULUA

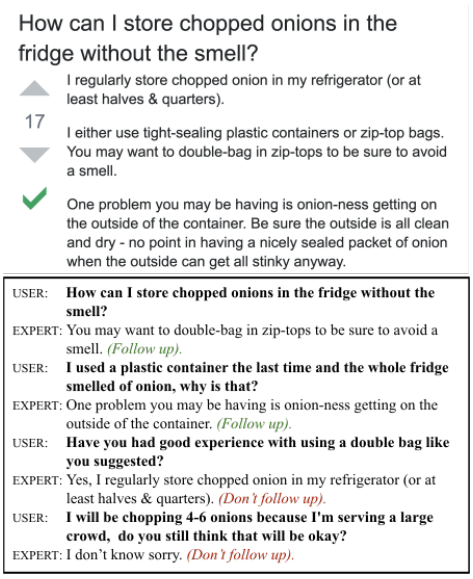
CQA datu multzoa

Kapitulu honetan, sortu dugun galdera-erantzunetan oinarritutako elkarrizketa datu multzoa aurkeztuko dugu. Bertan, hau sortzeko motibazioa, erabilitako metodologia eta teknikak aurkeztuko ditugu. Honekin batera datu multzoaren gainean egindako analisia ere azalduko da, artearen egoerako beste multzo batzuekin konparatuz.

3.1 Motibazioa eta ekarpenak

Informazio testuala bilaketa bidez atzitzeak bere mugak ditu. Orokorrean, *Google* bezalako bilatzaileek dokumentu osoak itzultzen dituzte kontsulta bakoitzerako eta honek erabiltzaileen eskuzko lana eskatzen du, dokumentu luze bat emanda bertan erantzuna topatu behar baitute. Azkenaldian, *chat* motako testu eta ahots bidezko informazio erauzketa geroz eta ezagunagoa bilakatzen ari da hau baita gizakien elkarrekintza modurik ohikoena. Honen ondorioz, erabiltzaileek erlazionatutako galderei modu motz eta zehatzean erantzuten dieten sistemak espero dituzte. Erantzuna osatugabea edo partzialki zuzena den kasuetan, erabiltzaileak jarraipen galderak egin ditzake honela, sistemari erabiltzailearen beharrak asetzeko aukera gehigarriak emanez.

Behar guzti hauek asetzeko garatu dugu DoQA datu multzoa. DoQA ohiko galderak atzitzeko (FAQ, ingelesezko *frequently asked question*-etik) galdera-erantzun motako elkarrizketaz osatua dago (CQA, ingelesezko *conversational question answering*-etik). Datu multzoak 2.437 elkarrizketa ditu hiru domeinu ezber-



3.1 Irudia – Sukaldaritzari buruzko elkarrizketa bat. Goian publikazio originala, gaia eta erantzun pasartea dituena. Honen azpian publikazio originalaren gaineko elkarrizketa ikus daiteke non erabiltzaile batek egindako galderak eta aditu batek erantzundako erantzunak ikus daitezkeen. Parentesi artean elkarrizketa ekintzak azaltzen dira, hauek adituak erabiltzen ditu erabiltzailea elkarrizketan zehar gidatzeko.

dinetan banatuta (10.917 galdera guztira). Elkarrizketa hauek *Wizard of Oz* teknika erabiliz sortzen dituzte bi *crowdworker*-ek.*Wizard of Oz* teknikan, subjektuek autonomoa dela uste duten sistema informatiko batekin elkarreragiten dute, baina sistema hau autonomoa izan ordez, gizaki batek operatzen du (Kelley 1985). Gure bi langileek hurrengo rolak hartzen dituzte: **erabiltzaileak** gai jakin bati buruzko galderak egiten ditu, eta **domeinuko adituak** galderei erantzuten die testu zati laburrak hautatuz erantzuna barne duen testu luze batean. Elkarrizketako lehen galdera jatorrizko FAQean definitutakoa izango da beti, eta galdera honek definituko du elkarrizketa gai nagusia ondorengo jarraipen galderentzat. Erantzuten duenak dokumentu luzean testu azpi-atal bat hautatzeaz gain, erantzuna birformulatzeko aukera ere badu, erantzun abstraktu eta naturalagoa emateko aukera ahalbidetuz. Datu multzoak erantzun ezin diren galderak eta elkarrizketa-ekintza garrantzitsu batzuk ere baditu. DoQAn bi azpi ataza desberdin definitzen ditugu: jatorrizkoan ebaluaziorako adibideak galdera eta helburu-dokumentuaz osatuak daude, non erantzunak topatu behar diren; informazioa berreskuratzeko (IR) esze-

natokian probako datuek galderak dituzte, baina helburu-dokumentua ezezaguna da, eta sistemak erantzunak dituzten dokumentuak hautatu behar ditu bildumako dokumentu guztien artean.

DoQAk hurrengo kontribuzioak egiten ditu:

- Aurretik sortutako irakurketa ulermen datu multzoak ez bezala erabiltzaileen benetazko informazio beharrak islatzen ditu, errealitatean existitzen den FAQ batean oinarrituta baitago. DoQAn lortzen diren emaitza positiboek erakutsiko lukete posible dela galdera-erantzun elkarrizketen bidez FAQetan dagoen informazioa modu eraginkor batean atzitzea.

- DoQAko elkarrizketa informazio behar errealetatik datozenez, datu multzoko elkarrizketak koherenteagoak, naturalagoak eta galdera konplexuagoz osatuak daude. Guzti hau datu multzoaren gainean egindako analisiek berretsi dute.

- DoQAn proposatutako informazio berreskurapen atazari eta datu multzoko domeinu desberdinei esker, aurretik garatutako datu multzoak baina erronka handiagoa da DoQA. Artearen egoerako sistemek lortutako emaitzek erakusten duten bezala, oraindik tarte handia dago makinen eta gizakien artean ataza honetan.

3.2 Metodologia

Atal honetan, datu multzoa biltzeko metodologia eta prozesua deskribatzen da. Prozesu honek, *Amazon Mechanical Turk*-en (AMT) bi *croworkdwer*entzat diseinatutako lan interaktiboa definitzen du.

3.2.1 Publikazioen aukeraketa

Ohiko galderak eskala handian erauzteko sarean atzigarri dauden datuak modu automatikoan deskargatu ditugu. Honetarako, *Stack Exchange* web-orri multzoa erabiltzen dugu, zeinean, domeinu desberdinetako galdera-erantzunak topa daitezken, hari bezala ere ezagutzen direnak. *Stack Exchange* plataformak domeinu desberdinetako galdera-erantzunak edukitzeak domeinu arteko transferentziako analisia egitea ahalbidetuko digu, bereziki datu urriko testuinguruetan fokua jarriko dugularik. Web-orri guztietatik *Stack Overflow*, *Super User* eta *Ask Ubuntu*

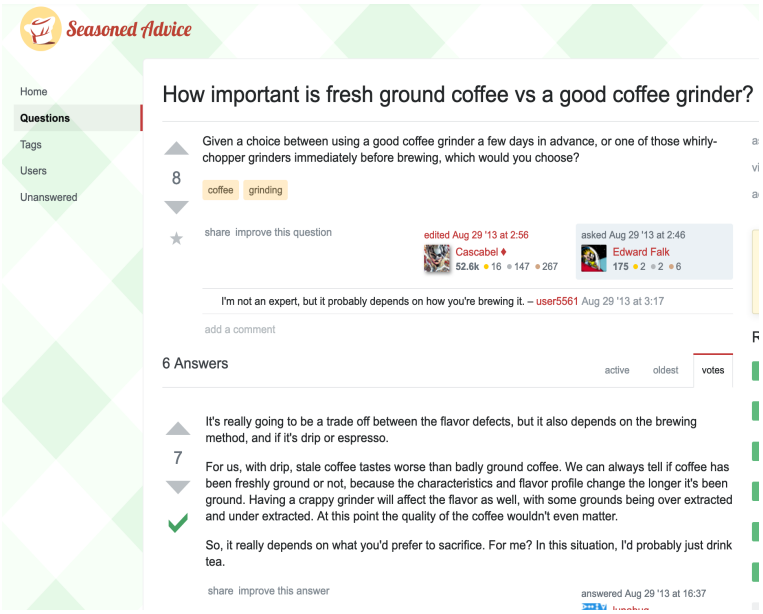
dira ezagunenak, zeinetan, kodeari, ordenagailuei eta *Ubuntu*-ri buruzko galdera-erantzun bikoteak topa ditzakegun hurrenez hurren. Orri hauen funtzionamendua hurrengoa da: erabiltzaileek galderak idazten dituzte domeinu zehatz batekin lotuak daudenak, ondoren beste erabiltzaile batek hauek erantzungo dituelarik. Galdera bakoitzerako erantzun bat baino gehiago egon daitezke, eta inongo aditurik ez dagoenez, edozein erabiltzailek erantzun bati aldeko edo aurkako botoa eman dakioke, erantzun onenak nabarmenduz. Mota honetako webguneak komunitate galdera-erantzun gune bezala ere ezagutzen dira.

Stack Exchange web-orriko datu-iraulketetatik hiru domeinu ezberdinetarako gai-erantzun bikoteak bildu ditugu. Azpi multzo guztietatik, sukaldaritzako [1], bidaietako [2] eta filmetako [3] domeinuetan zentratu gara. Domeinu hauek aukeratu ditugu, foro aktiboak direlako eta interes orokorreko ezagutza dutelako. Honela, datu hauen gainean elkarrizketak sortuko dituzten *crowdworker*entzat ulergarria eta erakargarria izango delarik. Kontuan izan, *StackExchange*ko argitalpenek (FAQ gehienetan bezala) galdera konplexuak dituztela, eta askotan erantzun luzeak eskatzen dituzte. Gainera, *community question answering* guneetan topatu daitezkeen erantzunak orokorrean modu aberatsean azalduak eta argudiatuak egoten dira, bereziki aldeko boto asko jaso dituztenak. Hau kontuan izanik, gure hipotesia hurrengoa da: *StackExchange*ko erantzun konplexu baten atzean galdera-erantzun motako elkarrizketa latente bat ezkutatzen da, non, azpi-galdera desberdinak definitu daitezkeen. *StackExchange* barruko edozein haritan hurrengo atalak nabarmendu ditzakegu:

- Hariko **galdera** nagusia, **izenburu** edo **gai** bezala ere ezagutzen dena. Hau haria irekitzeko idatzitako lehen galdera izango da, erantzuten duenaren rola hartzen duen erabiltzaileak erantzuten saiatu beharko dena. Plataformako beste edozein erabiltzailek galdera hauek balora ditzake, galderaren kalitatea ebaluatuz. 3.2 irudian adibidez jatorrizko galdera "*How important is fresh ground coffee vs a good coffee grinder?*" ("Zenbateko garrantzia du ehotutako kafe freskoak eta kafe-errotagailu on batek?") da eta 8 puntuko nota jaso du.

- **Galderaren aurrekariak**. *Stackexchange*n galdera berri bat idazten den bakoitzean, galdetzaileak honekin erlazazionatutako zenbait aurrekari eman ditzake, erantzungo duten pertsona potentzialek galdera hau hobeto ulertu eta kokatzeko. Testu hau beti izenburuaren atzetik agertzen da eta 3.2

[1] `https://cooking.stackexchange.com/`
[2] `https://travel.stackexchange.com/`
[3] `https://movies.stackexchange.com/`



3.2 Irudia – *Sukaldaritzako hari baten adibidea StackExchange*n. Irudiaren goiko aldean hariko galdera nagusia ikus daiteke. Galderaren azpian agertzen den testua galderaren aurrekari bezala ezagutzen da eta galdera erantzuteko beharrezkoa den informazioa gehituko du. Irudiaren behe aldean, galderari emandako erantzuna ikus daiteke.

irudian honela definitzen da "*Given a choice between using a good coffee grinder a few days...*" ("Egun batzuk kafe-errotagailu on bat erabiltzearen aukera emanda...").

- Erantzuten dutenek foroan emandako **erantzunak**. Hauek jatorrizko galderari erantzuna ematen saiatzen dira modu argi eta hedatuan ahalik eta balorazio onenak lortzeko beste erabiltzaileengandik. 3.2 Irudiko adibidean jatorrizko galderari 6 erantzun ematen zaizkio eta baloraziorik onenari beste erabiltzaileek 7 boto positibo eman dizkiote. Honetaz gain, jatorrizko galdera argitaratu duenak erantzunetako bat zuzentzat marka dezake, berarentzat baliagarriena zein izan den erakusteko.

Gure datu multzoa sortzeko 2018 iraileko *StackExchange*ko datu iraulketa deskargatu dugu, zeinak 19.818 hari dituen guztira. Puntu honetan galderen eta erantzunen luzeren analisi txiki bat egin dugu. Bertan ikusi ahal izan dugu nola galdera guztien puntuazioa [-6, 240] tartean dagoela. Ausaz galdera batzuk eskuz azter-

tu ostean, ikusi dugu nola puntuazio baxuko galderek ere kalitate altua duten, puntuazio negatibokoak izan ezik. Pasarteen luzera maximoa 2960 hitzekoa da, beraz, badaude pasarte batzuk luzeegiak direla gure atazarako. Azken finean, pasarte luzeek elkarrizketen sorrera asko zailduko dute anotatzaileentzat. Hau guztia kontutan hartuta, hurrengo iragaziak erabili ditugu *StackExchange*ko harietan:

- Puntuazio negatiboa duten galderak ezabatu ditugu ez baititugu kalitatea txarreko galderak nahi gure datu multzoan.
- Galdera ikur bat baino gehiago dituzten harien izenburuko galderak baztertu ditugu. Honen arrazoia da gure elkarrizketen lehen galdera hariko izenburua izango dela eta ez dugula nahi galdera bat baino gehiagoko txandarik elkarrizketetan.
- Pasarteen luzera 50 eta 250 hitz artean mugatzen dugu, hau baina luzeago edo motzago diren pasarteak baztertuz. Muga hauekin lortzen dugu pasarteak motzegiak ez izatea eta eduki minimo bat izatea baina luzeegiak izan gabe ataza anotatzaileentzat gehiegi ez zailtzeko.
- Hiperestekak, irudiak edo kodea bezalako HTML *tag*-ak dituzten hariak ezabatzen ditugu.

3.2.2 *Crowdsourcing* ataza

Anotazio prozesurako *Amazon Mechanical Turk* (AMT) plataforman *Human Intelligence Task* (HIT) bezala ezagutzen diren zenbait ataza definitu ditugu. AMT sareko *crowdsourcing* merkatu bat da zeinean zenbait tresna eskaintzen diren plataformako langileak eta HITak argitaratzen dituzten bezeroak koordinatzeko. HIT batek zeregin bakar, autonomo eta birtual bat adierazten du, *crowdworker* batek landu, erantzun bat bidali eta osatzeagatik ordainsari bat jasoko duena.

Plataforma honen funtzionatzeko modua hurrengoa da: lehenik eta behin bezeroek HIT kopuru mugatu bat argitaratuko dute osatuak izateko. Hau egitean bezeroak zenbait parametro zehaztu behar ditu. Lehenik, HIT bakoitzeko ordainsaria definitu behar da $ 0.01tik hasita. Ondoren, HITak ikusgai izango dituzten langileak iragazteko zenbait parametro zehaztu daitezke: langilearen HIT onarpen-tasa historiko minimoa, langileari onartutako HIT kopuru minimoa edo honen kokapena. Honetaz gain, proposatutako HITak atazaren aurretiko ezagutza eskatzen badu, bezeroak kualifikazio ariketa bat definitu dezake langileei bertan

puntuazio minimo bat lortzea eskatuz. Behin hau guztia zehaztu eta gero eginkizuna AMTn argitaratuko da eta langileen hau atzitu dezakete. Pauso honetan, zehaztutako iragazki guztiak gainditzen dituzten langileek baloratu dezakete ea ataza interesgarria egiten zaien eta honetan lan egin nahi dutenetz. Behin HIT guztiak osatuak izan direnean edo bezeroak ataza bertan behera uzten duenean bezeroak jasotako emaitzak onartu edo baztertu beharko ditu ondoren langileen zehaztutako diru kopurua ordaintzeko. Honetaz gain, *bonus* bat ordaintzeko aukera ere badago egindako lana nabarmena izan denean.

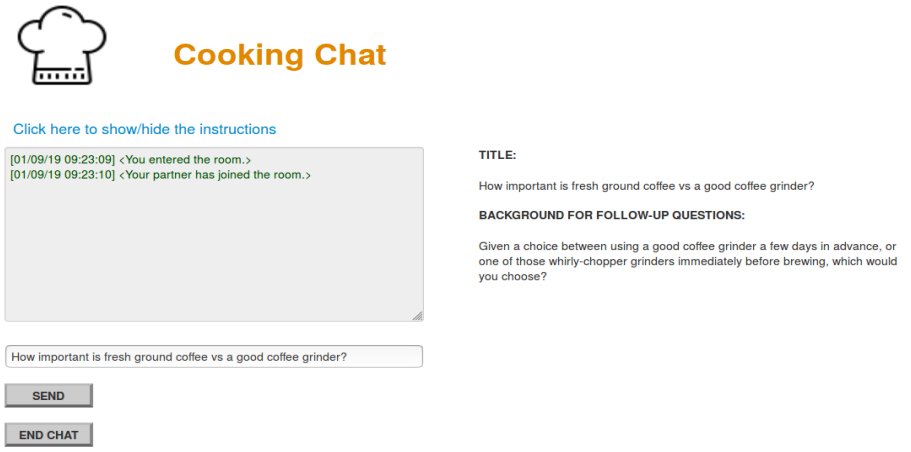
Funtzionamendu guzti honen arazo nagusia langileen babes falta da. Bezero ez zintzo batek langileen emaitzak baztertu ditzake ordainsaririk ez emateko nahiz eta hauen lana ona izan. Arazo honi aurre egiteko langileak foro desberdinetan [4] antolatzen dira, non eguneko HITen *feedback*a ematen duten.

Ataza interaktiboa Gure kasuan HIT bakoitza bi langileen artean gai zehatz baten inguruan elkarrizketa bat sortzeko ataza izango da. Langileetako batek, erabiltzailearen rola hartzen duenak, gai zehatz baten inguruko galderak egingo dizkio bigarren langileari. Bigarren langile honek adituaren rola hartuko du. Galdera guztiak *StackExchange*ko sukaldaritza, bidai edo filmetako harietatik erauziak izan dira.

Erabiltzailearen rola hartu duen langileak hariko galdera nagusia eta gaia aurkezten duen paragrafo txiki batera atzipena izango du. Informazio hau edukirik, erabiltzaileak edozein luzerako galderak egin beharko ditu. Dialogo guztietan lehen galdera *StackExchange*ko harian agertzen den gaiaren izenburua izango da. 3.3 Irudian erabiltzailearen interfazearen adibide bat aurkezten dugu. Interfazearen eskuin aldean hariaren izenburua eta gaiaren aurkezpen paragrafoa ikus daiteke. Ezkerraldean, ordea, orain arteko elkarrizketa azaltzen duen testu kutxa bat dago, galderak idazteko sarrera kutxa batekin batera. Honetaz gain, galdera bidali eta elkarrizketa amaitzeko botoiak ere inplementatu ditugu erabiltzailearentzat.

Domeinuko adituak erantzuna duen pasarte osoa ikusiko du eta erabiltzaileak egindako galdera-erantzun beharko du pasarte honetako azpi-zati bat hautatuz. Elkarrizketak itxura naturalagoa izan dezaten, adituak erantzuna moldatzeko aukera izango du, hala ere, garrantzitsua da kontuan hartzea behin erantzuna moldatzean ez dela bat etorriko jatorrizkoarekin. Hortaz, (Yatskar 2019) lana jarraituz, aldaketa minimoak motibatzen ditugu sortutako sareko interfazeko erantzun kutxan erantzuna zuzenean kopiatuz. Erantzuna itzultzeaz gain, adituak erabiltzailea gi-

[4] `https://turkerview.com/` and `https://turkopticon.ucsd.edu/` for example

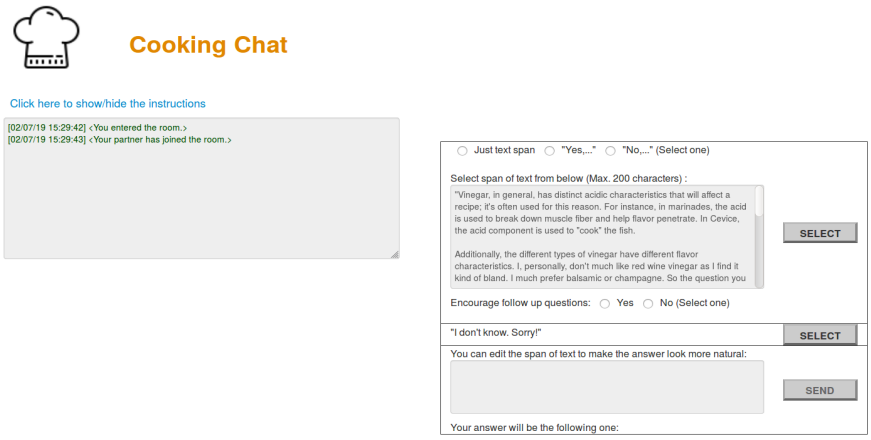


3.3 Irudia – Elkarrizketak biltzeko erabiltzailearen interfazea. Ezker aldean elkarrizketaren testuingurua ikus daiteke eta galdera idazteko testu kutxa. Eskuin aldean *StackExchange*ko hariaren izenburua eta gaia aurkezten duen paragrafoa.

datu beharko du elkarrizketan zehar zenbait elkarrizketa ekintza erabiliz. datu multzo hau sortzeko erabiltzen ditugu elkarrizketa ekintzat hurrengoak dira:

- Baieztapen ekintza, galdera bai/ez motakoa denean beharrezkoa dena (**bai**, **ez**, **bat ere ez**).
- Jarraipen ekintza, erabiltzaileak azpigai emankorrenekin erlazionatutako galderak egiten jarraitu dezan (**jarraitu**, **ez jarraitu**)
- Erantzungarritasun ekintza, adituak duen testuan galdera-erantzungarria den edo ez definituko duena (**erantzuna du**, **ez du erantzunik**)

Adituak erantzunik ez dagoela anotatzen duenean, itzultzen den erantzuna "*I don't know*" ("Ez dakit") karaktere katea izango da. Azpimarratzekoa da, jarraipen elkarrizketa ekintzaren helburua erabiltzailea galderak idazteko garaian modu lausoan gidatzea dela, honela adituak azken galderarekin lotuak dauden galderak egitera animatu edo desanimatu dezake erabiltzailea. Erabilitako elkarrizketa ekintzak QuACen definitutakoen berdinak dira, baina jarraipen ekintzatik agian jarraitzeko aukera kenduta, izan ere, egindako probetan ez zen intuitiboa anotatzaileentzat. 3.4 Irudian adituaren interfazearen adibide bat ikusi daiteke. Interfazearen eskuin aldean erantzuna aukeratzeko inplementazio ikusi daiteke. Ezkerrean, erabiltzailearen kasuan bezala, elkarrizketa azaltzen den testu kutxa bat



3.4 Irudia – Elkarrizketak biltzeko adituaren interfazea. Ezker aldean elkarrizketaren testuingurua ikus daiteke. Eskuin aldean, ordea, elkarrizketa ekintzak aurkezten dira, erantzuna aukeratu beharreko pasartearekin batera.

ikus daiteke. Kasu honetan ez dugu elkarrizketa amaitzeko botoia inplementatu ez baitugu nahi adituak elkarrizketa amaitzeko aukera izatea.

Elkarrizketa guztiak automatikoki amaitzen dira 8 galdera-erantzun bikotera iristean, erantzunik gabeko 3 galdera daudenean edo 10 minutuko kontagailua amaitzen denean. Muga hauen helburua elkarrizketa luze eta errepikakorrak saihestea da, *StachExchange*n aukeratutako domeinuen benetako hariak gai jakin bati oso bideratuta daudelako. Honetaz gain, ez da 2 galdera-erantzun pare baina gutxiago edo "Ez dakit"motako erantzunak bakarrik dituen elkarrizketarik onartuko.

Garatutako datu bilketa interfazea *CoCoa*n [5] (He *et al.* 2017) oinarritzen da. *CoCoa* elkarrizketak modu kolaboratiboan biltzeko *framework* bat da.

AMTn erabilitako parametroak AMTko langileak aukeratzeko hurrengo parametroak definitu ditugu HITak sortzean:

- HIT onarpen tasa $\geq 98\%$
- Onartutako HIT kopurua ≥ 1000
- Langileen kokapena: ingelesa hitz egiten den herrialdeak.

[5] `https://github.com/stanfordnlp/cocoa`

	Sukaldaritza			Bidaiak	Filmak
	Entrenamendua	Garapena	Test	Test	Test
Galderak	4,612	911	1,797	1,713	1,884
Elkarrizketak	1,037	200	400	400	400
Hariak	546	162	400	400	400
Galdera / Tokenak	10.79	10.14	10.66	10.45	9.45
Erantzun / Tokenak	13.19	13.10	12.58	13.47	12.40
Elkarrizketa txandak	4.47	4.55	4.49	4.28	4.71
Estraktibo %	69.68	67.18	66.95	65.44	74.15
Abstraktibo %	30.32	32.82	33.05	34.56	25.85
Bai/Ez %	20.22	21.07	22.20	25.10	18.05
Ez dakit %	27.55	27.33	29.71	22.83	29.41

3.1 Taula – DoQA datu multzoko estatistikak. Bertan ikus daiteke nola azpi multzo desberdinen ezaugarriak oso antzekoak diren nahiz eta domeinu desberdinetako datuak dituzten.

Langile bakoitzari 0,10 $ ordaintzen dizkiogu HITa egiteagatik eta 0,33 $-eko bonus bat sortutako galdera eta erantzun bakoitzeko, betiere erantzun hau "Ez dakit"motakoa ez bada. Ordainsarien distribuzio honek langileak erantzun zuzena topatzera motibatzen ditu, azken finean "Ez dakit"erantzutea askoz errazagoa da, ez baitago erantzuna duen pasartea irakurri beharrik. Elkarrizketa bakoitzaren batazbesteko prezioa 3,2 $-etakoa da.

3.2.3 Datu multzoaren xehetasunak

Ohiko praktika jarraituz, sukaldaritzako datu multzo nagusia entrenamendu, garapen eta test zatietan banatu dugu. Beste bi domeinuetarako, Bidaiak eta Filmak, test banaketa baino ez dugu. 3.1 Taulan agertzen dira domeinu eta banaketa guztien estatistikak.

Sukaldaritzako datu multzoaren banaketek oso antzeko ezaugarriak dituzte, beraz, banaketa guztiek distribuzio berdina jarraitzen dutela espero dezakegu. Test-banaketetan ez dugu hari berdinari buruzko elkarrizketa bat baino gehiago onartzen, test garaian ereduen orokortze gaitasuna ahalik eta hoberen ebaluatzeko.

3.5 Irudia – Erantzun anitzak biltzeko interfazea. Goiko aldean elkarrizketa testuingurua eta pasartea erakusten dira. Beheko aldean, ordea, anotatzaileak elkarrizketa ekintza aukeratzeko eta erantzuna idazteko testu kutxa azaltzen dira.

3.2.4 Erantzun ugari biltzen

DoQA datu multzoan gizakien errendimendua kalkulatu ahal izateko domeinu guztietako test multzoetan erantzun berriak bildu ditugu bigarren anotazio txanda batean. Erantzun anitzak biltzea ebaluaziorako ere lagungarria da, izan ere, DoQAko galderak erantzun posible bat baina gehiago izan ditzakete.

Bigarren anotazio txanda honetan, AMTko langileek lehen txandako galderak ikusiko dituzte, testuinguruko galdera-erantzunekin batera, eta erantzun bat aukeratu beharko dute jatorrizko adituaren rola imitatuz. Erantzun anitzak biltzeko interfazea 3.5 Irudian ikus daiteke.

3.3 Analisia

	DoQA	QuAC	CoQA
Galderak	10,917	98,407	127,000
Elkarrizketak	2,437	13,594	8,399
Galdera / Tokenak	10.43	6.5	5.5
Erantzun / Tokenak	12.99	14.6	2.7
Elkarrizketa txandak	4.48	7.2	15.2
Estraktibo %	69.13	100	66.8
Abstraktibo %	30.87	-	33.2
Bai/Ez %	21.01	25.8	-
Ez dakit %	27.47	20.3	1.3

3.2 Taula – DoQAren estatistikak QuAC eta CoQArekin konparatuta. DoQAren tamaina hiruetatik txikiena da baina galdera eta erantzun luzeenak ditu. CoQA ,ordea, nahiz eta handiena den tamainaz, ia galdera guztiak erantzungarriak ditu.

Estatistika orokorrak Atal honetan DoQA multzoaren analisi kuantitatibo eta kualitatibo bat aurkezten da. Analisi honetan, DoQA aurretik garatutako antzeko datu multzoekin konparatzen dugu, antzekotasunak eta desberdintasunak nabarmenduz. Analisirako kontutan hartzen ditugun datu multzoak galdera-erantzun motako elkarrizketaz osatutako QuAC eta CoQA dira.

3.2 Taulan DoQAren estatistika orokorrak azaltzen dira, QuAC eta CoQA-koekin batera. Bertan ikus daitekeen bezala, DoQA galdera eta dialogo gutxien dituen datu multzo da, hala ere, honek dituen beste ezaugarri guztiek oso interesgarria egiten dute galdera-erantzun motako elkarrizketak ikertzeko. Galdera eta erantzun bakoitzeko batazbesteko token kopurua (10.43 eta 12.99, hurrenez hurren), adibidez, elkarrizketa naturalek dituztenetatik gertuago dago. Azken finean, elkarrizketa esanguratsua bada, parte-hartzaile bakoitzak bere ikuspuntua modu informagarrian erakutsi nahiko du eta honetarako txanda luzeak beharko ditu. CoQAren kasuan adibidez, ikus daiteke nola oso galdera eta erantzun motzez osatua dagoen, galdera-erantzun sinpleez osatutako datu multzo bat dela iradokiz. Elkarrizketa bakoitzak duen galdera kopurua aztertuz ikus daiteke DoQA dela proportzio txikiena duena, izan ere, DoQA hari bakar bat jorratzen elkarrizketaz osatua dago. Behin erabiltzaileak bere beharra ase ostean, honek elkarrizketa amaituko du, elkarrizketa luze eta korapilatsurik sortu gabe. Azkenik aipatu, CoQAko galdera gehienek erantzun bat dutela, SQuAD 1.0 (Rajpurkar *et al.* 2016) datu multzoak dituen arazo antzekoak erakutsiz. SQuADeko autoreek azaldu zu-

ten bezala, galdera guztiak erantzungarriak izatea arazo larri bat da ereduek beti erantzun bat topatuko baitute, zer dakiten eta zer ez dakiten ikasi gabe. Arazo hau izan zen hain zuzen SQuAD 2.0 (Rajpurkar *et al.* 2018) datu multzoan erantzungarriak ez ziren galderak gehitzeko motibazioa.

Honetaz gain, datu multzoa sortzeko garaian, AMTko langileentzat inkesta txiki bat diseinatu dugu datuen kalitatea neurtu ahal izateko. Elkarrizketa bakoitza sortu eta gero, erabiltzailearen rola duen langileak jasotako erantzunak ebaluatu behar ditu 1-5 *likert* eskala batean. Erabiltzaileen batez besteko asetzea 3,9koa da DoQAn. Bestalde, adituaren rola duen erabiltzaileak jasotako galderen kalitatea eta emandako erantzunen lagungarritasuna neurtu behar ditu. Ataza honetan lortutako batazbesteko balioak 4,27 eta 4,10 izan dira, AMTko ataza egokia izan dela erakutsiz.

Naturaltasuna Gure datu multzoaren alde positiboetako bat elkarrizketen naturaltasunean oinarritzen da, izan ere, ezaugarri hau ez da QuAC bezalako antzeko datu multzoetan ikusten. DoQAko erantzunak sarean atzigarri dauden foroetatik datoz eta web-orri hauetan dagoen testua pertsona bati zuzendua da, jatorrizko galdera idatzi zuen erabiltzaileari hain zuzen ere. QuAC eta CoQAko erantzunak, ordea, *Wikipedia*tik datoz eta bertan dagoen testua askoz ere testu formalagoa da. Formaltasun hau ez dago elkarrizketetan erabiltzen den erregistrotik gertu, hortaz, foroetako estiloa askoz ere egokiagoa da elkarrizketa baterako. 3.1 Irudiko elkarrizketa orain aipatutako naturaltasunaren adibide garbi bat da, non adituak erabiltzaileari zuzenean egiten dion erreferentzia "*You may want* "("Baliteke zuk nahi izatea") edo "*you may be having* "("zuk izan dezakezula") esaldien bitartez. DoQA-ko elkarrizketak QuAC-ekoak baina naturalagoak direla erakusteko 50 elkarrizketa ausaz lagindu ditugu eta A/B test bat aurrera eraman dugu proiektuan parte ez den anotatzaile batekin. Ataza honetan anotatzaileak DoQAko eta QuACeko elkarrizketa bat ikusirik bietatik naturalagoa zein iruditzen zaion aukeratu behar du. Anotazio prozesu honek erakutsi du DoQAko elkarrizketak kasuen % 84 batean naturalagoak direla.

Naturaltasun honen jatorria da DoQAren elkarrizketak behar zehatz bat duen erabiltzaile batek hasten dituela. Honen ondorioz, jarraipen galderak oso erlazionatuak daude aurreko erantzunekin eta galdera guztiek hari berdina jarraitzen dute. QuACen kasuan, ordea, elkarrizketek ez dute helburu argi bat aurkezten eta galderak ausaz galdetuak dirudite. Gainera, DoQAko elkarrizketak jatorrizko informazio beharra asetzean amaitzen dira, elkarrizketa natural batean espero genukeen bezala.

Laginaren analisi sakonago batek erakutsi du DoQAko erantzunak espontaneoagoak direla ahozkotasunaren berezko ezaugarriak erakusten baitituzte. Hauen artean adierazkortasun maila altuagoa: "*Normally when I try they end up burned not crispy!*" ("Normalean probatzen dudanean erre egiten zaizkit eta ez zaizkit kurruskarri gelditzen!"), "*My biggest worry here would be...*" ("Nire kezkarik handiena hemen ... izango litzateke"), "*hey let's not be hasty* "("Ez gaitezen presaka ibili"), iritziak :"*I came across a suggestion to cover the lid...*" ("Tapa estaltzeko iradokizun bat topatu nuen"), "*I'd recommend simply adding...*" ("Besterik gabe, ... gehitzea gomendatuko nuke) eta humorea: "*well yeah but booze is booze*" ("beno bai baina alkohol gogorra alkohol gogorra da") nabarmendu ditzakegu. Bestalde, QuACeko erantzunak hermetikoagoak dira eta ez dute elkarrizketa bateko ahozkotasuna erakusten. Ezaugarri guzti hauek DoQAko elkarrizketak naturalagoak bilakatzen dituzte.

Honetaz gain, QuAC DoQA baina naturalagoa zen elkarrizketen % 16a sakonago aztertu dugu eta ikusi dugu nola dialogo hauetan galderak ez ziren modu zuzenean erantzuten. Hurrengo galdera (G) eta erantzun (E) pareek honen adibide bat erakusten dute: (G) "*Is the taste going to be significantly different?*" ("Zaporea nabarmen ezberdina izango al da?") (E) "*there is cornstarch in confectioner's sugar*" ("azukrean arto-almidoia dago"); (G) "*how about reheating?*" ("Eta berriro berotzen saiatzen bazara?")(E) "*When you defrost it, do so in your fridge leaving it overnight so that it defrosts gradually*" ("Desisoztu ezazu hozkailuan gau osoan utziz, pixkanaka desizoztu dadin"); (G) "*Can I use my potatoes or carrots if they already have some roots?*" ("Erabili al ditzaket nire patatak edo azenarioak dagoeneko sustrai batzuk badituzte?") (E) "*The green portions of a potato are toxic*" ("Patataren zati berdeak toxikoak dira"). Elkarrizketa hauetako batzuetan erantzun zuzena ez da adituak atzigarri duen pasartearen parte. Kasu hauetan, adituak erantzun oker bat eman beharrean "Ez dakit"erantzun beharko luke.

Galdera motak 3.3 Taulak sukaldaritzako datu multzoko galderetako hasierako bi hitz usuenak erakusten ditu, hauen agerpen portzentaiekin batera. Galdera gehienak *what* ("zer") eta *how* ("nola") hitzekin hasten dira (galderen % 16.6 eta % 15.1 hurrenez hurren) QuAC eta CoQA datu multzoen kasuan bezala. Hala ere, sukaldaritzako datu multzoko galderak ezin dira erantzun sinple batekin erantzun "How long"("zenbat") kasuan izan ezik. Bestalde, CoQA eta QuACen kasuan, galderen hasierako hitz gehienek (*who*, *where*, *when* ("nor", "non"eta "noiz"hurrenez hurren)) datu zehatz batekin erantzun daitezkeen galderak direla erakusten dute. Hipotesi hau berresteko, 50 ausazko galdera eskuz aztertu ditugu

Bigrama aurrizkia		%	Adibidea
What	is	30.8	What is the purpose of adding water to an egg wash?
(16.6%)	are	8.0	What are other methods to sharpen a knife?
How	do	24.0	How do you properly defrost frozen fish?
(15.1%)	long	21.9	How long should I cook it in the microwave?
Is	there	52.8	Is there a special tool available for cracking open a pistachio?
(10.5%)	it	19.8	Is it safe to cook with rainwater?
Do	you	70.7	Do you have any advice for storing green onions?
(7.6%)	I	16.1	Do I have to peel the apples?
Can	I	52.8	Can I put them back in the oven to reheat?
(5.5%)	you	25.3	Can you explain the science behind this cooking procedure?
I	have	19.6	I have been told that frying it would make it tastier, but is it healthier to grill or fry?
(5.0%)	am	15.3	I am cooking for somebody who doesn't eat shellfish, so is the fish sauce safe?
Why	is	22.1	Why is it important to increase the fermentation time?
(3.5%)	does	21.7	Why does my custard pudding taste like raw eggs?

3.3 Taula – Sukaldaritza domeinuko lehen hitz eta bigrama usuenak. Adibideetan ikus daiteke nola galdera usuenak konplexuak diren eta ezin diren erantzun sinple batekin ebatzi.

eta egiaztatu dugu DoQAko galderen % 66ak erantzun konplexuak behar dituela sukaldaritza domeinuan. Honek erakusten du DoQAko galdera gehienak irekiak direla. Analisi berdina QuAC egiterako garaian galdera konplexuen portzentaia %36-ra erortzen da. Emaitza hauek ez datoz guztiz bat (Choi *et al.* 2018) lanean azaltzen dituztenekin, bertan galderen erdiak konplexuak direla esaten baitute.

Testuinguruaren menpekotasuna DoQAko galderak testuinguruarekiko menpekotasuna dutenetz azaltzeko bigarren eskuzko analisi bat egin dugu. Analisi honetan berriro ere 50 galdera ausaz lagindu ditugu eta hauek testuinguruarekiko menpekotasuna duten edo ez anotatu dugu. Bertan, galderen % 61 batek koreferentzia bezalako fenomenoak dituela azaldu da. Adibidez, "*What are other methods to sharpen a knife?"* ("Zeintzuk dira labana zorrozteko beste metodo batzuk?"), "*How long should I cook it in the microwave?"* ("Noiz arte egosi behar dut mikrouhin labean?"), "*Can you explain the science behind this cooking procedure?"* ("Azal al dezakezu sukaldaritza prozedura honen atzean dagoen zientzia?").

Elkarrizketen koherentzia Elkarrizketa bakoitzak koherentzia bat jarraitzen duela egiaztatzeko A/B test berri gauzatu dugu DoQA eta QuACen gainean. A/B test honetan DoQA-ko sukaldaritza domeinuko eta QuACeko 50 elkarrizketa la-

	DoQA	QuAC	CoQA
Benetazko informazio beharra	✓		
Naturaltasuna	✓		
Elkarrizketen koherentzia	✓		
Galdera konplexuak	✓		
Erantzunik gabeko galderak	✓	✓	
Elkarrizketa egintza	✓	✓	
Domeinu anitzak	✓		✓
Informazio Berreskurapen ataza	✓		

3.4 Taula – DoQA datu multzoaren ezaugarrien laburpena QuAC eta CoQArekin konparatuz.✓ezaugarria betetzen duen kasurako erabiltzen dugu.

gindu ditugu eta bietatik koherenteena dena aukeratu du anotatzaileak, hau da, elkarrizketa fluxu leunagoa duena. Anotazio ariketa honek erakutsi du DoQAko dialogoak QuACekoak baina koherentegoak direla kasuen % 64 batean. Aurrakoa kasuen % 10 batean bakarrik gertatzen da eta kasuen beste % 26a berdinketa bat da. DoQAk galtzen duen kasuen %10 hori aztertu dugu eta ikusi dugu nola elkarrizketek kasu hauetan oso antzeko galderak dituzten, bata bestearen atzetik, errepikakorrak bihurtuz eta koherentzia galduz.

3.4 Taulak aipatutako DoQAren ezaugarri positiboak laburtzen ditu QuAC eta CoQArekin konparatuz.

4. KAPITULUA

CQA datu multzoaren gaineko esperimentuak

Kapitulu honetan, DoQA datu multzoaren gainean egindako oinarrizko esperimentuak aurkezten ditugu, domeinu arteko transferentzian foku berezia jarriz. Kapituluak egitura berdina jarraitzen duten bi atal nagusi ditu. Atal hauetan esperimentuen helburua, jarraitutako metodoa, oinarrizko ereduak, konfigurazio esperimentala eta emaitzak azaltzen ditugu. Lehenik makina bidezko irakurketa ulermen atazan (MRC, ingelesezko *machine reading comprehension*-etik) egindako esperimentuak aurkezten dira eta ondoren informazio berreskurapenekoak.

4.1 Domeinu arteko transferentzia

Helburua

Esperimentu hauen helburua CQA atazan artearen egoerako sistemek duten domeinu arteko transferentzia gaitasuna aztertzea da. Honetarako domeinu orokorreko QuAC datu multzoa eta domeinu zehatzak dituen DoQA datu multzoak erabili ditugu.

Atazaren deskribapena

Domeinu arteko transferentzia ebaluatzeko ez dugu elkarrizketa testuingurua kontuan hartuko, beraz, elkarrizketa motako galdera-erantzun sistema ataza honela formalizatzen dugu kasu honetan: p erantzuna duen pasartea eta q_k galdera eman-

	Sukaldaritza		Bidaiak		Filmak	
Konfigurazioa	F1	HEQ-Q	F1	HEQ-Q	F1	HEQ-Q
Jatorrizkoa	40,1	35,1	36,2	34,8	36,1	33,5
Zero-shot	40,2	34,7	34,0	30,1	38,2	33,2
Transferentzia	**43,3**	**37,8**	**40,6**	**33,6**	41,8	**36,3**
Transferentzia osoa	43,1	37,0	40,6	33,4	**42,0**	34,5
Pertsona	86,6	100,0	87,4	100,0	88,8	100,0

4.1 Taula – BERT oinarrizko ereduaren emaitzak DoQAko hiru domeinuetan (zutabeak) lau konfigurazio desberdinetan (ilarak). Irakurri testua ilaren azalpenerako. Kontuan izan filmen eta bidaien emaitzak domeinu barruko inongo entrenamendu daturik gabe lortu direla.

da, sistemak a_k erantzuna itzuli beharko du, p pasarteko i indizean hasi eta j indizean amaitzen dena.

Erabilitako sistemak

Domeinu arteko transferentzia aztertzerko tesiko atal hau garatu zen garaiko artearen egoerako sistema bat erabili genuen. Sistema hau aurrentrenatua izan den BERT (Devlin *et al.* 2019) ereduaren galdera-erantzun sistema bezala erabiltzeko aldaera bat da. Aldaera honek, p pasartea eta q_k galdera emanik, a_k erantzunaren i hasiera eta j amaiera indizeak aurrikusiko ditu. Honetarako, bi bektore berri gehitzen dira doikuntza garaian, hasiera bektore bat $S \in \mathbb{R}^H$ eta amaiera bektore bat $E \in \mathbb{R}^H$. $T_i \in \mathbb{R}^H$ p pasarteko i indizea duen hitzaren azken errepresentazio ezkutua izanik, hitz hau erantzunaren hasiera izateko probabilitatea $Ps_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$ erabiliz kalkulatuko da. Formula berdina erabiltzen da amaiera probabilitateak kalkulatzeko. Azkenik, erantzun zuzena aukeratzeko $a_k = argmax_{i,j} Ps_i + Pe_j$ erabiliko da. Sistema doitzeko jatorrizko (Devlin *et al.* 2019) artikuluan proposatzen diren doikuntza hiperparametroak erabili ditugu. Egindako esperimentu guztietan 110M parametro dituen BERTen *base-uncased* aldaera erabiltzen dugu.

Emaitzak

Esperimentu hau ebaluatzeko DoQAko informazio estraktiboa bakarrik erabili dugu, berridazketa abstraktiboak etorkizunerako lan bezala utziz. Domeinu transfe-

rentzia ebaluatzeko hurrengo konfigurazioa zehazten dugu:

- **Jatorrizko konfigurazioan** DoQAko sukaldaritzako entrenamendu eta garapen datuak erabiltzen ditugu, lehengoan sistema entrenatzen dugu bigarrenean ebaluatzen dugun bitartean. Hobekuntzarik ikusten ez dugun unean entrenamendua eteten dugu. Teknika hau goiz-eten bezala ezagutzen da.

- **Zero-shot** kasuan QuACen entrenamendu eta garapen datu multzoak erabiltzen ditugu entrenatzeko eta goiz-etena aurrera eramateko.

- **Transferentzia konfigurazioan** DoQAko sukaldaritza domeinua eta QuAC erabiltzen ditugu entrenatzeko.

- **Transferentzia osoa** eszenatokian QuAC eta DoQAko sukaldaritza domeinuaz gain ebaluatzeko erabiltzen ari ez garen beste domeinuetako ebaluazio multzoak ere erabiltzen ditugu entrenatzeko.

4.1 Taulak emaitza guztiak laburtzen ditu. Bertan ikus daiteke nola metrika guztiek (metrika desberdinen azalpena 2.1.2 atalean ikus daiteke) antzeko joera duten, beraz, F1ean zentratuko gara emaitzen eztabaida errazago aurrera eramateko.

Kasu desberdinetan arreta jarriz, lehenik eta behin *Sukaldaritza* domeinuan zentratuko gara. Jatorrizko eta *zero-shot* konfigurazioetan oso emaitza antzekoak lortzen ditugu. Honek erakusten du nola Sukaldaritzako 1000 eta Wikipediako 13000 elkarrizketek errendimendu bera erakusten duten. Garrantzitsua da aipatzea QuAC DoQAko tamaina berdinera ausaz azpilagintzen badugu sukaldaritza domeinuko emaitzak 36,5ra erortzen direla. Bi datu multzoak, DoQA eta QuAC, konbinatzean emaitzak 7 puntu hobetzen dira ("Transferentzia" zutabea) eta oraindik hobekuntza handiagoak ikus daitezke filmen eta bidaien domeinuen datuak gehitzean doikuntza gehigarrirako ("Transferentzia osoa" zutabea).

Hala eta guztiz ere, gure esperimentuetan ikusten den joera interesgarriena bidaien eta filmen domeinuetakoa da, izan ere, esperimentu hauetan ez dago inongo domeinu barruko entrenamendu daturik. Kasu honetan, jatorrizko eta transferentziako domeinuek ez dute domeinu barruko entrenamendurik eta sukaldaritzako emaitzekin konparagarriak dira ($\sim$ 3 F1eko okertzea). Beraz, esan daiteke ez dela beharrezkoa domeinu bakoitzerako entrenatu behar izatea maiz egiten diren galderen domeinuan eta beste domeinuetako FAQen entrenamendu datuak oso berrerabilgarriak direla.

Ondorioak

Domeinuz kanpoko (bidaiak eta filmak) ebaluazio datu multzoko elkarrizketetan lortutako emaitzak deigarriak dira *Wikipedia* eta sukaldaritzako datuetan entrenatzean, izan ere, domeinu barruko datuekin bakarrik lortutako emaitzen oso gertu daude sukaldaritzako ebaluaziorako elkarrizketetan. Gure hipotesia da erabiltzaileek *Stackexchange* motako FAQ web-orrietan ezaugarri linguistiko antzekoak erabiltzen dituztela, 3.2 atalean azaldutakoekin erlazionatuak. Bestalde, *Wikipedia*ko testuak helburu desberdin batekin sortu dira eta ezaugarri linguistiko desberdinak erakutsiko dituzte. Adibide bezala, FAQetan ez bezala, *Wikipedia*ko testuak ez ditu lehen eta bigarren pertsonako izenordainak. Analisi linguistiko sakonagoa egin liteke datu multzoaren gainean baina hau ez da tesi honen helburu nagusietako bat, beraz, etorkizuneko lan bezala definitzen da.

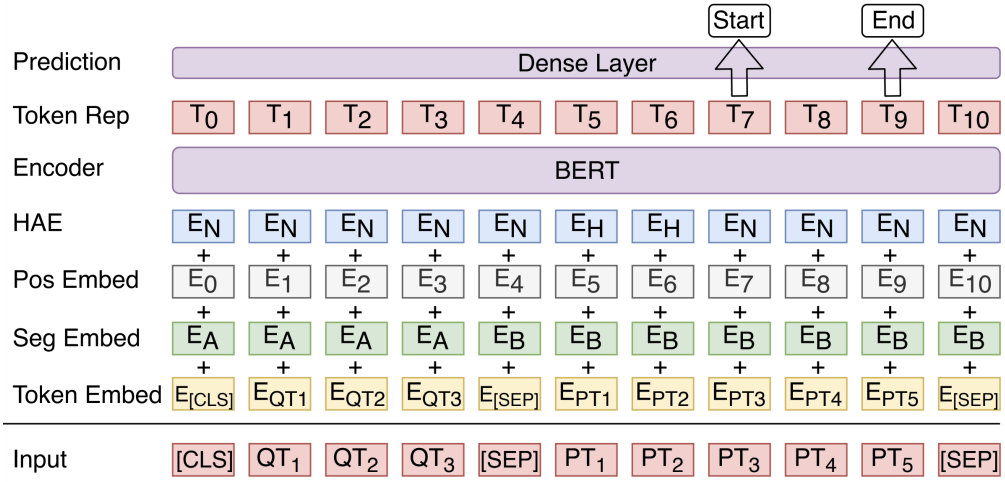
4.2 Testuinguruaren eragina

Helburua

Bigarren esperimentu multzo honen helburua DoQA datu multzoaren testuinguru dependentzia aztertzea da. CQA datu multzo batek elkarrizketa txanda bakoitzean testuinguruaren dependentzia izan beharko luke eta uneko galdera ondo erantzuteko testuingurua modu egokian prozesatzea ezinbestekoa izan behar da. Hau honela ez balitz QA datu multzo baten aurrean egongo ginateke eta hau ez da gure garapenaren helburua. Hau erakusteko aurreko ataleko esperimentuak errepikatzen ditugu baina testuingurua kontuan hartzen duen sistema bat erabiliz.

Atazaren deskribapena

CQA ataza honela formalizatzen dugu: p erantzuna duen pasartea eta elkarrizketaren testuingurua emanda $\{q_1, a_1, ..., q_{k-1}, a_{k-1}\}$, non testuinguruko galderak $q_1, .., q_{k-1}$ eta datu multzoko erantzun zuzenak $a_1, .., a_{k-1}$ dauden, sistemak a_k erantzuna itzuli beharko du, p pasarteko i indizean hasi eta j indizean amaitzen dena. Honetaz gain, v elkarrizketa ekintzen lista ere aurrikusteko aukera dago. Elkarrizketa ekintzen listak *{yes,no,-}* etiketak ditu baieztapena aurrikusteko eta *{follow-up,don't follow-up}* jarraipena aurrikusteko. Hala ere, esperimentu hauetan ez dugu v aurrikusten ez baitu testuinguruaren erabilera aztertzeko balio.



4.1 Irudia – BERT+HAE sistemaren errepresentazioa. Bertan ikus daiteke nola HAE errepresentazioak sistemako azpi-hitzei gehitzen zaizkien hauek aurretik testuinguruan agertu badira. Irudiaren jatorria: (Qu *et al.* 2019)

Erabilitako sistemak

Testuinguruaren erabilera aztertzeko, elkarrizketako aurreko txandak kontuan hartzen dituen sistema bat erabiltzen dugu (Qu *et al.* 2019). Sistema hau aurreko atalean erabilitako BERT **base uncased**en gainean eraikitzen dugu. BERTen aldaera honek elkarrizketaren testuinguruko erantzunen menpeko hitz bektoreak (HAE ingelesezko *History Answer Embedding*-etik) (Qu *et al.*, 2019) erabiltzen ditu. Hortaz, elkarrizketa testuinguruaren informazioa $\{q_1, a_1, ...q_{k-1}, a_{k-1}\}$ gehitzen dio BERT eredu bati testuinguruko erantzunen hitz bektoreen geruza bat gehituz, zeinean p pasarteko token bakoitza testuinguruaren parte denetz ikasiko den. Ondoren errepresentazio hauek BERT sistemako azpi-hitz errepresentazioei gehituko zaizkie. Honen errepresentazio bisual bat ikus daiteke 4.1 Irudian.

Emaitzak

Esperimentu hauetan aurretik erabilitako konfigurazio berdinak erabiltzen ditugu baina BERT ereduari HAE errepresentazioak gehituz. Gogoan izan erabiltzen ditugun konfigurazioak hurrengoak direla: jatorrizkoa, *zero-shot*, transferentziazkoa eta transferentzia osokoa.

4.2 Taulan emaitza guztiak ikusi ahal dira. Aurreko kasuan bezala metrika guztiek joera antzekoa dutenez, F1ean zentratuko gara emaitzen eztabaida erraza-

	Sukaldaritza		Bidaiak		Filmak	
Konfigurazioa	F1	HEQ-Q	F1	HEQ-Q	F1	HEQ-Q
Jatorrizkoa	47,8	43,0	44,0	37,4	42,8	37,1
Zero-shot	46,2	42,0	42,7	37,1	45,4	41,4
Transferentzia	53,2	**48,3**	50,8	42,1	51,6	44,3
Transferentzia osoa	**53,4**	46,9	**51,6**	**43,3**	**52,1**	**45,2**
Pertsona	86,6	100,0	87,4	100,0	88,8	100,0

4.2 Taula – Testuingurua kontuan hartzen duen BERT+HAE sistemaren emaitzak DoQAko hiru domeinuetan (zutabeak) lau konfigurazio desberdinetan (ilarak). Kontuan izan filmen eta bidaien emaitzak domeinu barruko inongo entrenamendu daturik gabe lortu direla.

go aurrera eramateko. Konfigurazio eta domeinu guztietan BERT+HAE ereduek emaitza hobeak lortzen dituzte aurreko ataleko BERTek baino, argi erakutsiz DoQA datu multzoan elkarrizketaren testuingurua garrantzitsua dela eta testuinguruko galderak eta erantzunak modelatzea ezinbestekoa dela.

Garrantzitsua da aipatzea QuAC ausaz DoQAko tamaina berdinera azpilagintzen badugu sukaldaritza domeinuko emaitzak 36,5ra erortzen direla. Bi datu multzoak, DoQA eta QuAC, konbinatzean emaitzak 7 puntu hobetzen dira ("Transferentzia" zutabea) eta oraindik hobekuntza handiagoak ikusi daitezke filmen eta bidaien domeinuen datuak gehitzean doikuntza gehigarrirako ("Transferentzia osoa" zutabea). Garrantzitsua da azpimarratzea sukaldaritza domeinuan lortutako errendimendua "Transferentzia" eta "Transferentzia osoa" konfigurazioetan konparagarria dela QuACen artikuluan aurkezten diren emaitzekin, kasu guzti hauetan entrenamendu eta test datuak domeinu berdinetik baitatoz. Nahiz eta kasu guzti hauetan entrenamendua eta testa domeinu berdinetik etorri, emaitzak DoQAn QuACen baina 9 puntu baxuagoak dira (BERT+HAEk 62,4eko F1a lortzen du), DoQA datu multzoa egungo hizkuntza ereduentzat erronka handiagoa dela erakutsiz.

Ondorioak

Esperimentu multzo honetako ondorio nagusia DoQA datu multzoan testuinguruak garrantzi handia duela da. Honetaz gain, azpimarratzekoa da nola DoQA QuAC datu multzoa baina erronka handiagoa den aurrentrenatutako hizkuntza

ereduentzat, datu multzoaren garrantzia bermatuz.

4.3 Domeinu irekiko galdera-erantzun sistema

Helburua

Tradizionalki irakurketa ulermen motako galdera-erantzun atazetan pasarte zuzena jasotzen du sistemak. Eszenatoki errealista batean, ordea, hau ez da gertatzen, galdera bat jasotzean sistemak ez baitu jakingo erantzuna zein pasartetan dagoen. Ataza errealistagoa egiteko intentzioarekin, domeinu irekiko galdera-erantzun (ODQA, ingelezko *open domain question answering*-etik) aldaera bat proposatu dugu DoQAren gainean, non erantzuna duen pasartea berreskuratu beharko den lehen IR pauso batean.

Atazaren deskribapena

Ataza hau bi pausotan banatzen da: lehenik IR sistema batek D dokumentu multzo batetik dokumentu potentzialak berreskuratu beharko ditu; bigarren pausoan, behin dokumentu hautagaiak ditugunean, QA sistemak erantzuna topatuko du. Erabiltzaile batek galdera bat egiten duenean, sistemak aurretik erantzun diren galdera antzekoak bilatu ditzake Dn hauek jasotako erantzuna pasarte bezala erabiltzeko. Beste aukera bat, zuzenean pasarteen artean bilatzea izan daiteke. Hortaz, bi metodo desberdin definitzen ditugu modu automatikoan erabiltzaileak egindako galderaren erantzuna duen pasartea topatzeko:

- **Galderen berreskurapena**, non hautagien artean galdera antzekoak bilatzen diren eta hauen erantzuna hartzen den pasarte potentzial bezala.
- **Erantzun berreskurapena**, non erantzun potentzialak zuzenean bilatzen diren kandidatu guztien artean.

IR ataza aurrera eramateko 20 pasarte esanguratsuenak berreskuratzen ditugu aurretik azaldutako bi metodoentzat. Honetarako bi indize desberdin sortzen ditugu, bat galderen berreskurapenerako eta bestea erantzunen berreskurapenerako. Lehenengo kasurako *StackExchange*n argitaratutako gaiak indexatzen ditugu. Bigarrenean, ordea, *StackExchange*ko pasarteak indexatzen ditugu zuzenean. Metodo hau DoQAko garapen eta ebaluazio azpi multzoetako elkarrizketetan aplikatzen dugu, elkarrizketako lehen galdera erabiliz berreskurapen galdera bezala.

Kontuan izanik datu multzoko elkarrizketak gai bakar baten inguruan direla, elkarrizketako lehen galdera bakarrik erabiltzen dugu informazioa berreskuratzeko eta lortutako pasartea elkarrizketako galdera guztiak erantzuteko erabiliko dugu.

Behin pasarte hautagiak ditugunean QA sistema aplikatzen dugu hautagai hauen gainean. Sailkapen honek 20 pasarte berreskuratzen ditu elkarrizketa bakoitzeko, honek esan nahi du, elkarrizketa bakoitzeko 20 hautagai izango ditugula. Honi aurre egiteko hurrengo hiru aukerak ebaluatzen ditugu:

- Lehen esperimentuan, *Top-1*, *IR*k itzulitako sailkapeneko lehen pasartea bakarrik hartzen dugu kontuan.

- Bigarren esperimentuan, *Top-20:BERT*, pasarte guztiak ematen dizkiogu pasarte posible bezala BERT hizkuntza ereduari eta probabilitate altueneko erantzuna aukeratzen dugu zuzentzat. Kontutan izan, "Ez dakit" erantzunak dituzten pasarteak baztertzen ditugula.

- Hirugarren esperimentu batean, *Top-20:BERT*IR*, pasartea aukeratzeko *IR* motorrak itzulitako puntuazioa BERT sistemaren probabilitateekin konbinatzen dugu eta balio maximoa duena hautatzen dugu.

Erabilitako sistemak

Informazio berreskurapen atazarako errepresentazio sakabanatuak erabiltzen dituen IR sistema publiko bat erabiltzen dugu. Sistema hau *BM25* algoritmoan oinarritzen da (ikusi 2.1.2 Atala azalpen sakonago batentzat). Bigarren pausoan, testuingurua kontuan hartzen ez duen BERT eredua erabiltzen dugu. Kasu guztietan defektuzko hiperparametroak erabiltzen ditugu.

Emaitzak

4.3 Taulak galdera eta erantzun berreskurapen metodoen emaitzak azaltzen ditu. Galdera berreskurapen metodoak oso emaitza onak lortzen ditu, 0,94ko errendimendua lortuz P@1 metrikan. Hau espero zitekeen emaitza bat da, izan ere, AMTko langileek hariko galdera nagusia atzitu dezakete, askotan hau bera erabiliz lehen galdera bezala. Nahiz eta kasu batzuetan edizio txiki batzuk gehitzen dituzten, datu multzoko sarrerako galdera gehienak IR sistemak indexatuta dituenen oso antzekoak dira. Erantzun berreskurapenaren kasuan, ordea, emaitzak txarragoak dira, 0,54ko marka lortuz P@1 metrikan.

Berreskurapena	MAP	P@1	R@20
Galdera	0,95	0,94	0,98
Erantzuna	0,65	0,54	0,88

4.3 Taula – Galdera eta erantzun berreskurapen emaitzak sukaldaritza domeinurako. Batezbesteko doitasuna (MAP, ingelesezko *mean average precision*-etik), 1eko doitasuna 1ean (P@1, ingelesezko *precission*-etik) eta 20ko estaldura (R@20, ingelesezko *recall*-etik) metrikak aurkezten ditugu.

Eredua	F1	HEQ-Q
Erantzun berreskurapena		
Top-1	**37,2**	**33,3**
Top-20:BERT	32,7	29,6
Top-20:BERT*IR	36,1	32,9
Galdera berreskurapena		
Top-1	**42,2**	**36,76**
Top-20:BERT	35,8	31,2
Top-20:BERT*IR	41,6	36,4

4.4 Taula – Sukaldaritza domeinuko emaitzak ODQA eszenatokian. Ikusi testua azalpenentzat.

4.4 Taulak ODQA eszenatokiko emaitzak aurkezten ditu. Bertan, hurbilpen sinpleenak, *Top-1*, lortzen ditu emaitza onenak bai galdera eta erantzun berreskurapen estrategietan. Informazio berreskurapeneko teknika garatuagoak aztertzea tesi honen markotik haratago doa eta arrazoi horregatik ez da azterketa sakonagorik egin. Galderen berreskurapena erabiliz lortzen diren emaitzak 4.1 Taulan azaldu ditugunetatik oso gertu daude. 2.1.2 atalean aurkeztutako erantzun berreskurapen emaitzak kontuan izanik, harrigarria da 5 puntuetako galera bakarrik ikustea galdera eta erantzun berreskurapen metodoen artean. Ikusi ahal izan dugunagatik, korrelazio handia dago elkarrizketa sistemen eta erantzun berreskurapen sistemen erroreen artean eta honek desberdintasun txiki hau azalduko luke. Bi berreskurapen estrategietan emaitzak gertu daude pasarte zuzena atzitzea posible den kasuetatik.

Ondorioak

Esperimentu hauekin ikusi dugu nola ataza errealistagoa egiten dugunean sistemen kalitatea txarragoa den, hala ere, sistemek duten galera ez da espero bezain handia IR sistemarentzat zailak diren adibideak QA sistemarentzat ere hala direlako.

5. KAPITULUA

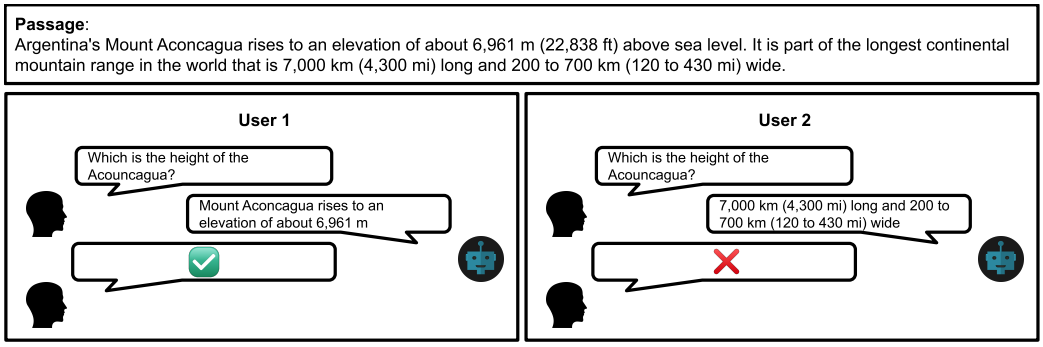
Erabiltzaile *feedback* bitarra

Atal honetan aurretik entrenatua izan den galdera-erantzun sistema bat jarriko dugu gizaki simulatuekin elkarrekintzan. Honekin batera, metodo berri bat aurkeztuko dugu galdera-erantzun sistemaren errendimendua hobetzeko elkarrekintza hori probestuz. Egiturari dagokionez, lehenik, lan hau motibatuko dugu ondoren erabilitako metodologia eta emaitzak aurkezteko. Azkenik, metodoaren mugak azalduko ditugu, honen gainean garatutako aldaerak motibatuko dituztenak.

5.1 Motibazioa eta ekarpenak

Aurreko kapituluan erakutsi dugun bezala, CQA sistemak gai dira gizakiekin galdera-erantzun motako elkarrizketak mantentzeko. Sistema hauek DoQA bezalako datu multzoetan entrenatzean errendimendu ona dutela erakutsi dugu, baina datu multzo hauek gizaki-gizaki elkarrekintzen bitartez sortzen dira eta hau oso prozesu garestia da.

Elkarrizketa sistemak modu naturalean elkarreragiten dute gizakiekin eta honek aukera asko irekitzen ditu behin sistemak martxan jarri eta gero hobetzen jarraitzeko. Datu multzo estatiko bat izanda enpresa batek elkarrizketa sistema on bat garatu dezake eta martxan jarri. Behin martxan jarri ostean, erabiltzaileen elkarreraginetik informazio aberasgarria lortzeko aukera egon daiteke sistema eguneratzen eta hobetzen jarraitu ahal izateko. Erabiltzaileetatik jaso daitezkeen seinale desberdinetatik kapitulu honetan *feedback* bitar esplizituan jartzen dugu fokua, *feedback* esplizituetatik sinpleena baita. Ataza orokorraren adibide bat 5.1



5.1 Irudia – Galdera-erantzunetan oinarritutako sistema baten adibide bat, non erabiltzaileak jasotzen duen erantzun bakoitzeko sistemari *feedback* bitarra emango dion. Jatorria: (Campos *et al.* 2020).

Irudian ikus daiteke. Bertan galdera-erantzun motako elkarrizketa bat dugu makina eta pertsona baten artean. Galderez gain, kasu honetan erabiltzaileak *feedback* bitarra emango dio makinari eta makinak hau erabili behar du informazio honi esker hobetzen jarraitzeko.

Kapitulu honetan hurrengo ekarpenak egiten ditugu:

- Sistema bat martxan jarri eta gero, erabiltzaile *feedback* bitarra jasoz sistemaren eguneratzea gaitzen duen algoritmo baten garapena.
- Dokumentu sailkapen eta CQA atazetan garatutako esperimentuetan sistema gainbegiratu bat *feedback* bitarra bakarrik erabiliz hobetzea posible dela erakustea.
- Sistema martxan jarri eta gero domeinu aldaketa bat gertatzen denean, *feedback* bitarra aldaketa horretara egokitzeko informazio seinale aberatsa dela erakustea.

5.2 Metodologia

Sistema martxan jarri eta gero hobetzeko eszenatokian lehenik eta behin jatorrizko S_0 sistema bat entrenatzen dugu modu gainbegiratuan. S_0 sistema honek ikaskuntza automatikoko lan-fluxu tradizionala jarraitzen du, non entrenamendu eta garapen datu multzo gainbegiratu mugatu batera atzipena dugun. Entrenamendu fase honen ondoren, garapenean emaitza onenak lortzen dituen sistema martxan

jartzen dugu, erabiltzaileen galderak ebatzi ditzan. Bigarren fase honetan, gizaki batek x galdera bat egiten duen bakoitzean, sistemak y erantzun bat sortzen du eta gizakiak *feedback* bitarra ematen du honen gainean. Denbora aurrera doan ahala, sistemak erantzun desberdinak sortzen ditu $y_{i1}, y_{i2}, ..., y_{in}$ eta *feedback* bitarra jasoko du x_i elementu bakoitzarentzat. Gure esperimentu guztietan gizaki makina elkarrekintza nahikoa izango ditugula suposatzen dugu, hortaz, ez dugu *feedback* gabeko adibiderik kontuan hartzen. Sistema gizakiekin elkarreraginean egon eta gero, gizakien galderak, sistemaren erantzunak eta gizakien *feedback*ak jasotzen ditugu.

3. kapituluan bezala, kasu honetan ere galdera-erantzun sistema bi sailkatzaileren bitartez inplementatzen dugu. Sailkatzaile hauetako batek hasiera tokena aurreikusiko du eta besteak, ordea, bukaera tokena. Horri esker, sailkatzaile bakoitza modu independentean tratatu dezakegu. Sistemaren erantzunak sortzeko, autonormalizatutako garrantzi bidezko laginketan oinarritzen den *feedback*-pisudun ikaskuntza erabiltzea proposatzen dugu.

5.2.1 Ikasketa algoritmoa

Atal honetan, diseinatutako ikasketa algoritmoa definitzen dugu. Algoritmo honi esker, datu multzo estatiko batean entrenatutako sailkatzaile bat eguneratu dezakegu modu dinamiko batean gizakien *feedback*a bakarrik erabiliz. Lehenik eta behin, C klaseen gaineko eta x sarrera duen gure helburu distribuzioa $p^*(y|x)$ definitzen dugu. Distribuzio hau *feedback* bitarra islatzeko sortzen dugu $\{-\beta, \beta\}$:

$$p^*(y|x) \propto \begin{cases} exp(\beta), & y \text{ zuzena bada} \\ exp(-\beta), & y \text{ zuzena ez bada} \end{cases}$$

Hitzetan, klase bakoitzaren zuzentasuna klaseari esleitutako probabilitatearen magnitudean islatzen da, erabiltzaileen iritziarekiko proportzionala dena. β hiperparametroak *feedback*aren pisua kontrolatzen du. Proposatutako algoritmoaren helburua da p^* eta gure sailkatzailearen $q(y|x;\theta)$ arteko KL dibergentzia minimizatzea, θ hiperparametroekiko, non

$$\text{KL}(p^*|q) = -\sum_y p^*(y|x) \log q(y|x;\theta) + \mathcal{H}(p^*). \tag{5.1}$$

Helburu funtzio hau minimizatzea konputa ezina da $y$ren espazioa esponentziala delako. Honen ondorioz, ezinezkoa izango da $p^*(y|x)$ren normalizazio

konstantea konputatzea. Arazo hau ebazteko *Monte Carlo* laginketa hurbilpena erabili dugu, $p^*(y|x)$ distribuziotik laginak sortuz. Soluzio honek, ordea, arazo bat du oraindik, ez baitugu $p^*(y|x)$ atzitzeko aukerarik. Hortaz, garrantzi bidezko laginketa bezala ezagutzen den teknika erabiliko dugu. Teknika honi esker p^* distribuzioko probabilitate ez normalizatuak lortu ditzakegu x sarrera bat eta y klase hautagai bat emanik. Garrantzi bidezko laginketa aurrera eramateko hurrengo proposamen banaketa definitzen dugu:

$$\hat{q}(y|x) = \lambda q(y|x;\theta) + (1-\lambda)\mathcal{U}(y), \tag{5.2}$$

non $\mathcal{U}(y)$ distribuzioa $y$rekiko distribuzio uniformea den eta q sailkatzailearen distribuzioa leuntzen lagunduko duen. Leunketa neurtzeko λ (Hoi *et al.* 2018) hiperparametroa definitzen dugu, honek esplorazioa eta esplotazioa neurtzen lagunduko digu. $\hat{q}(y|x)$ proposamen distribuzioarekin hurrengo helburu funtzioa definitzen dugu (5.1) Ekuaziotik hasita. Helburu funtzio hau izango da *feedback*-pisudun ikasketa bezala definitu dugun helburu funtzioa.

$$\begin{aligned}\mathrm{KL}(p^*|q) - \underbrace{\mathcal{H}(p^*)}_{\text{const. }\theta\text{rekiko}} &= -\sum_y \hat{q}(y|x)\frac{p^*(y|x)}{\hat{q}(y|x)}\log q(y|x;\theta) \\ &\approx -\frac{1}{K}\sum_{k=1}^{K}\frac{\omega(y^k)}{\sum_{k=1}^{K}\omega(y^k)}\log q(y^k|x;\theta),\end{aligned} \tag{5.3}$$

non K guztira jasotako *feedback* kopurua den eta pisuen autonormalizazioa erabiltzen dugun. $\omega(y^k)$ garrantzi bidezko laginketako pisuak honela kalkulatuko ditugu:

$$\log\omega(y^k) = \underbrace{\beta\mathbf{1}(y^k = y^*)}_{=\text{feedback}} - \log\hat{q}(y|x), \tag{5.4}$$

non y^* klase zuzen ezezaguna den eta

$$\mathbf{1}(\alpha) = \begin{cases} 1, \ \alpha \text{ zuzena bada} \\ -1, \ \alpha \text{ zuzena ez bada} \end{cases}$$

Beste hitz batzuekin, garrantzi bidezko laginketako pisuek erabiltzaileen *feedback*aren eta ereduaren konfiantzaren arteko proportzioa erakusten dute lagineko y^k iragarpen bakoitzeko. Hortaz, algoritmo hau *feedback*-pisudun ikasketa (FWL, ingelesezko *feedback weighted learning*-etik) bezala izendatzen dugu.

Algorithm 1: Erabiltzaile *feedback*a simulatzen datu multzo estatiko bati esker

Input: $\hat{a}_i$, $a_{i,j}$
Output: F_i (i erantzunari emandako *feedback*a)
$F1 \rightarrow F1(a_{i,j}, \hat{a}_i)$;
if $F1 = 1$ **then**
  Feedback positiboa eman(F_i);
else
  Feedback negatiboa eman(F_i);
end

5.2.2 Erabiltzaileen simulazioa

Erabiltzaile *feedback*a informazio iturri merkea da sistema errealak martxan jartzerakoan, hau naturalki jaso baitaiteke gizakien interakzioari esker. Esperimentuak garatzeko, ordea, guztiz kontrakoa gertatzen da, esperimentu bakoitzean *feedback*a uneko sistemaren menpekoa baita. Honek esan nahi du sistema zehatz bat erabiliz *feedback* datu multzo estatiko bat biltzean posible dela hau lagungarria ez izatea beste sistema batzuetarako, sistemak egiten dituzten akatsak desberdinak izan daitezkeelako.

Hau guztia dela eta, erabiltzaile *feedback*a datu multzo estatiko batekin simulatzeko oinarrizko algoritmo bat definitzen dugu. 1 Algoritmoan ikusi daiteke *feedback*a simulatzeko jarraitutako metodoa. Bertan, sistemak lagindutako $a_{i,j}$ erantzuna eta datu multzoko $\hat{a}_i$ anotazioak $F1$ maximoa badute, hau da, lagina eta erantzun zuzenak berdinak badira, *feedback* positiboa jasoko du sistemak eta kontrako kasuan negatiboa.

5.3 Esperimentuak

Atal honetan *feedback*-pisudun ikasketa algoritmoa erabiliz egindako esperimentuak aurkeztuko ditugu. Hauek aurkezteko aurreko kapituluko formatu bera jarraituko dugu. Esperimentu bakoitzean honen helburua, atazaren deskribapena, erabilitako sistemak eta emaitzak azalduko ditugu.

Esperimentu hauek aurrera eramateko lehenik eta behin S_0 izeneko jatorrizko sistema gainbegiratu bat sortzen dugu. S_0 jatorrizko sistema hau izango da hain zuzen ere gizakiekin martxan jarri eta geroko fasea simulatzeko erabiltzen dugu-

na. Simulazio honetan, sistema erreal baten joera erreplikatuko dugu S_0 sistemak sortutako erantzunei erabiltzaile simulatuen *feedback*a emanez. Erabiltzaile simulatuak eskuz anotatutako datu multzo estatiko batetik eratorri ditugu, entrenamendu datu multzoa zatituz. Puntu honetatik aurrera, S_0 entrenatzeko erabili dugun datu multzoari **entrenamendu multzoa** deituko diogu eta erabiltzaileen *feedback*a simulatzeko erabili dugun azpi multzoari, ordea, **simulazio multzoa**.

Ondorengo sistemak zehazten ditugu gure esperimentuetan:

- S_0: jatorrizko sistema gainbegiratua, entrenamendu datu multzoan bakarrik entrenatua izan dena. Eredu hau oinarrizko sistematzat hartzen dugu.
- S_0 + *FWL*: S_0 FWL algoritmoa erabiliz birdoitzen dugu, simulazio azpimultzoko adibideak eta *feedback* bitarra erabiliz.
- S_0 + *gainbegiratua*: lehenik eta behin S_0 entrenatzen dugu entrenamendu multzoan eta ondoren simulazioko azpi multzoko adibideak modu gainbegiratuan erabiltzen ditugu birdoitze garaian. Bertan klase zuzena ezagutuko dugu *feedback* bitarra jaso ordez. Hortaz, hau benetazko datuen atzipena duen sistema bat izango da.
- *Guztiz gainbegiratua*: sistema gainbegiratu bat hutsetik entrenamendu eta simulazio multzoen datuak elkartuz entrenatua izan dena.

Nahiz eta gure helburua etengabeko ikasketa erabiltzen duen CQA sistema bat garatzea den, garapena dokumentuen sailkapen ataza batean egiten dugu. Garapen honi esker, FWL algoritmoaren sendotasuna erakutsi dezakegu arkitektura eta ataza desberdinetan. Honetaz gain, esperimentu hauek hiperparametro esplorazioa egiteko ere erabiltzen dira, hortaz, dokumentu sailkapeneko hiperparametro konbinazio onena erabiliko dugu ondoren CQA sistema garatzean.

5.3.1 Dokumentuen sailkapena

Helburua

Dokumentu sailkapen atazan egindako esperimentuen helburu nagusia FWL algoritmoaren hiperparametro esplorazioa egitea da. Nahiz eta ataza nagusia CQA den, ataza honetan erabiltzen diren sistemak konputu gaitasun handiagoa eskatzen dute eta ez da bideragarria esplorazioa sistema konplexu hauekin egitea *hardware* mugen ondorioz.

Atazaren deskribapena

Dokumentu sailkapen atazan, x dokumentu bat izanik eta C klase hautagia, x dokumentuaren y klase zuzena aurrikusi behar du sistemak. Kasu honetan *DBPedia Classes* datu multzoan erabili dugu [1], 342.748 Wikipedia artikuluen kategoria hierarkikoz osatua dagoena. Artikulu bakoitza hiru maila desberdinetan sailkatua dago 9, 70 eta 219 kategoria desberdin izanik maila bakoitzean. Aukera guztietatik azkena erabiltzen dugu gure esperimentuetan, hortaz, dokumentu bakoitzerako 219 kategoria posible ditugu, hau da, $C = 219$ izango da. Garapen eta test datu azpi multzoak bere hortan mantentzen ditugu, baina entrenamendu multzoa bitan banatzen dugu entrenamendu eta simulazio azpi multzo berriak sortzeko. Entrenamendu azpi multzo berriak jatorrizko datuen % 10 izango du eta simulazio azpi multzoak datuen % 90a. Adibideak ausaz aukeratu ditugu. Portzentaia hauek eszenatoki errealak simulatzeko definitzen ditugu, izan ere, orokorrean sistemek dituzten entrenamendu adibideak mugatuak izaten dira eta lortzeko garestiak, erabiltzaileekin martxan jarri eta gero, ordea, errazagoa izan daiteke datuak modu merkean lortzea.

Erabilitako sistemak

Dokumentu sailkapenerako erabiltzen dugun eredua geruza anitzeko perzeptroi (MLP, ingelesezko *multi layer perceptron*-etik) bat da. Erabiltzen dugun MLPak geruza ezkutu bakar bat izango du. Eredu honen sarrera dokumentu bektoreak izango dira, dokumentuko hitzen *GloVE* (Pennington *et al.* 2014) bektoreen batezbesteko bezala kalkulatzen direnak. Bektoreen dimentsionalitatea 300ekoa izango da eta ezkutuko geruzak 200 unitate izango ditu.

Emaitzak

Lehenik eta behin S_0 sistema entrenamendu multzoaren gainean entrenatzen dugu entropia gurutzatua erabiliz galera funtzio bezala. S_0 + *FWL* sistemarentzat $\lambda \in [0,5, 1, 0]$ eta $\beta \in [1, 85]$ hiperparametroen esplorazioa egiten dugu optimizazio bayestarra erabiliz (Snoek *et al.* 2012). *Epoch* baten ondoren, garapen multzoan emaitza onenak lortzen dituzten hiperparametroak aukeratzen ditugu optimo bezala, hauek $\lambda = 0, 97$ eta $\beta = 76$ direlarik. Sarrera adibide bakoitzerako 3 klase desberdin lagintzen ditugu, eta S_0 + *FWL* 50 *epoch*etan entrenatzen dugu. Atal honetan *epoch* bat 3 klase lagintzearen baliokidea da.

[1] `https://www.kaggle.com/danofer/dbpedia-classes`

Sistema	F1
S_0	86,51
S_0 + *FWL*	91,59 (+5,0)
S_0 + *gainbegiratua*	91,89 (+5,3)
Guztiz gainbegiratua	92,04 (+5,5)

5.1 Taula – Dokumentu sailkapeneko emaitzak. FWL algoritmoak S_0 hobetzen jarraitzen du *feedback* bitarra bakarrik erabiliz. Emaitzak sistema gainbegiratuetatik gertu daude.

5.1 Taulan ikus daiteke nola MLP arkitektura sinple batek oso emaitza onak lortzen dituen ataza honetan, nahiz eta entrenamendu datu guztien % 10 bakarrik erabili dugun. Hala eta guztiz ere, S_0 + *FWL* algoritmoak S_0ren errendimendua hobetzen du 8 puntu baino gehiagotan eta beste bi sistema gainbegiratuetatik gertu dago. Emaitza hauek FWL metodoaren eraginkortasuna erakusten dute sistema gainbegiratu bat hobetzeko feedback bitarra bakarrik erabiliz.

5.3.2 CQA ataza

Helburua

Esperimentu hauen helburua da FWL algoritmoa ebaluatzea CQA atazaren gainean. Ataza hau bereziki garrantzitsua da testuinguru honetan CQA jatorriz ataza interaktiboa baita. Bertan gizakiak eta makinak txanda ugari dituen elkarrizketa bat garatzen dute, non *feedback*a esplizitua edo inplizitua izan daitekeen. Nahiz eta esperimentu hauetan kasu esplizitua bakarrik aztertzen dugun esperimentu hauek oso baliagarriak dira *feedback*aren gaineko ikasketaren erreferentzia gisa.

Atazaren deskribapena

4. Kapituluan egin dugun antzera, CQA ataza ebaluatzeko hurrengo aldaerak definitu ditugu:

- Domeinuen arteko esperimentuak. Kasu honetan bi eszenatoki desberdin erabiltzen ditugu sistema bat entrenatzeko garaian eta hau martxan jartzean gerta daitekeen domeinu aldaketa kontuan hartzeko. Lehen kasuan, domeinu barrukoan, entrenamendu eta simulazio datuak domeinu beretik datoz.

Domeinu kanpoko esperimentuetan, ordea, simulazioko domeinua entrenamendukoarekiko desberdina izango da.

- Elkarrizketa testuinguruarekin erlazionatutako esperimentuak. Bigarren esperimentu multzo honetan elkarrizketa testuinguruaren eragina aztertuko dugu. Honetarako, uneko galderari aurreko galdera eta erantzuna kateatzen dizkiogu, (Qu *et al.* 2019) lanean egiten duten bezala.

Domeinu barruko esperimentuetan QuAC (Choi *et al.* 2018) datu multzoa erabiltzen dugu, bai entrenamendu eta simulaziorako. Domeinu kanpoko esperimentuetan, ordea, QuAC S_0 sortzeko erabiltzen dugu baina simulaziorako DoQA datu multzoa erabiltzen dugu. Dokumentu sailkapenean bezala, entrenamendu multzo originalak bi azpi multzotan banatzen ditugu, % 10 entrenamendurako eta % 90 simulaziorako. QuACeko ebaluazio datuak atzigarri ez daudenez, garapen multzoko emaitzak erakusten ditugu.

Sistemaren hiperparametroei dagokienez, S_0 sistemaren λ eta β hiperparametroetarako dokumentu sailkapeneko konbinazio onena erabiliko dugu ($\lambda = 0,97$ eta $\beta = 76$). Ataza hau dokumentu sailkapenekoa baina konplexuagoa denez eta klase kopurua ere handiagoa denez, galdera bakoitzeko lagin kopurua 3tik 50era handitzen dugu. Kontuan hartu, galdera-erantzunetarako doitutako BERT sistemak bi sailkatzaile izango dituela, hasiera eta bukaera indizeak aurrikusteko, beraz, bigarren ataza hau dokumentu sailkapena baina konplexuagoa da.

Erabilitako sistemak

Galdera-erantzun motako elkarrizketen gaineko esperimentuak garatzeko 3. kapituluko metodologia bera erabili dugu. Honetarako, BERTen (Devlin *et al.* 2019) galdera-erantzun atazarako aldaera erabiltzen dugu. Honek, galdera eta erantzuna duen pasarte bat emanda hasiera eta bukaera indizeak aurreikusiko ditu. Esperimentu hauetan *BERT base uncased* eredua erabiltzen dugu hiperparametro lehenetsiekin.

Emaitzak

5.2 taulan QuAC datu multzoko domeinu barruko esperimentuen emaitzak aurkezten dira. Sistema bakoitzean eredu hoberenaren emaitzak aurkezten ditugu. Bertan ikus daitekeen bezala, FWL S_0ren gainean aplikatzeak 2,6 eta 4 puntu hobetzen ditu, FWL teknikaren erabilgarritasuna erakutsiz CQA sistema bat doitzen jarraitzeko. Elkarrizketaren testuingurua erabiltzeak sistemaren emaitzak 3

Sistemak	Testuinguru gabe	Testuinguruarekin
S_0	46,76	49,03
S_0 + *FWL*	49,33 (+2,6)	53,07 (+4,0)
S_0 + *gainbegiratua*	53,66 (+6,9)	55,10 (+6,1)
Gainbegiratu osoa	54,50 (+7,7)	55,40 (+6,5)

5.2 Taula – Domeinu barneko esperimentuen emaitzak. Bertan, QuAC datu multzoa erabiltzen dugu entrenamendu eta simulaziorako, elkarrizketa testuinguruaren eragina ikusteko. F1 balioak QuACeko garapen azpi multzoan kalkulatzen ditugu. FWLk S_0 hobetzen du, honen erabilgarritasuna berretsiz.

puntuz hobetzen ditu, testuingurua modelatzearen garrantzia azpimarratuz. Garrantzitsua da azpimarratzea nola FWL sistema gainbegiratuaren errendimendutik gertu gelditzen den *feedback* bitarra bakarrik erabiliz.

5.3 Taulak domeinuen arteko esperimentuen emaitzak aurkezten ditu. Bertan, S_0 QuACen gainean entrenatzen da eta ondoren DoQA erabiltzen dugu simulazio datu multzo bezala. Kasu honetan eredu hautapena DoQAko sukaldaritza garapen multzoaren gainean egiten dugu, ondoren test datu multzoan ebaluatzen dugularik. DoQAko ebaluazioa hiru domeinuetan banatzen da: sukaldaritza, bidaiak eta filmak. Bertan elkarrizketaren testuingurua kontutan hartzen duten sistemen emaitzak bakarrik aurkezten ditugu, hauek baitira emaitza hoberenak ematen dituzten sistemak. S_0 + *FWL* sistemak, S_0 hobetzen du domeinu guztietan. Gainera, S_0 + *FWL* gai da S_0 + *gainbegiratua*ren mailan aritzeko sukaldaritza eta filmen domeinuetan. *Guztiz gainbegiratua* sistemak S_0+ *gainbegiratuak* baina emaitza txarragoak lortzen ditu datu multzo honetan. Honen arrazoietako bat QuACen tamainan egon daiteke, izan ere, hau DoQA baino askoz handiagoa baita, DoQAko adibide bakoitzeko QuACeko 3 adibide daudelarik. Tamaina desberdintasunaren ondorioz, guztiz gainbegiratutako sistemak QuAC datu multzoarekiko alborapena izango du, DoQAn emaitza txarragoak lortuz. Garrantzitsua da azpimarratzea S_0 + *gainbegiratua* sisteman QuAC datu multzoa entrenamendurako bakarrik erabiltzen dela, ondoren doikuntza DoQAn egiten dugularik. Honen ondorioz S_0 + *gainbegiratua*k emaitza hobeak lortzen ditu orokorrean. Hau guztia kontuan hartuta, emaitza hauek erakusten dute FWL algoritmo sendo bat dela domeinu aldaketa bat dagoenean entrenamendutik ebaluaziora.

Sistema	Sukaldaritza	Filmak	Bidaiak
S_0	39,79	40,89	35,64
S_0 + *FWL*	49,66 (+9,9)	47,28 (+6,4)	47,19 (+11,6)
S_0 + *gainbegiratua*	50,63 (+10,8)	46,79 (+5,9)	47,12 (+11,5)
Guztiz gainbegiratua	50,33 (+10,5)	45,56 (+4,7)	46,10 (+10,5)

5.3 Taula – Domeinu arteko esperimentuen emaitzak QuAC entrenamendurako erabiliz eta DoQA, ordea, simulaziorako. DoQAko sukaldaritza, film eta bidaien domeinuetako emaitzak aurkezten ditu taulak. FWLk S_0 hobetzen du eta eredu gainbegiratuaren maila lortzen du bi domeinuetan.

5.4 Ondorioak

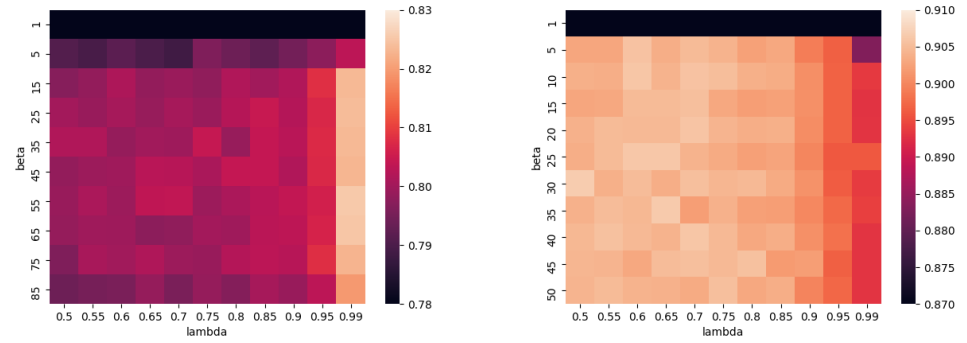
Garatutako dokumentu sailkapen eta CQA esperimentuetan erakutsi dugun bezala, gai gara S_0 jatorrizko sistema bat garatzeko erabiltzaileak simulatuz lortutako *feedback* bitarra bakarrik erabiliz. Esperimentu guztietatik bereziki interesgarriak dira entrenamendua eta gero domeinu aldaketa bat jasaten duten kasuak, hau sistema erreal askok topatzen duten zailtasun bat baita.

Hiperparametroak

FWL metodoaren sendotasuna erakusteko esperimentu asko garatu ditugu λ eta β hiperparametroen balio desberdin asko erabiliz. Analisi honek erakutsi du nola βri 1 baina handiagoak diren balioak ematean FWLk errendimendu antzekoa erakusten duen konbinazio desberdinetan (ikusi 5.2a eta 5.2b irudiak). Irudi berdinetan λren portaera aztertzen badugu ikusi daiteke nola lehen pausoetan λ handiak emaitza hobeak lortzen dituen, baina entrenamendua luzatzen dugun heinean λ txikiagoak gailentzen diren. Ohiko kasuetan ez bezala, FWLn esplotazioak entrenamenduaren hasieran laguntzen du eta esplorazioak entrenamendua aurrera joan ahala emaitzak hobetzen ditu. Honen arrazoietako bat FWLn aurretik doitutako sistema batetik hasten garela izan daiteke eta ez ausaz hasieratutako batekin.

Ikasketa dinamikak

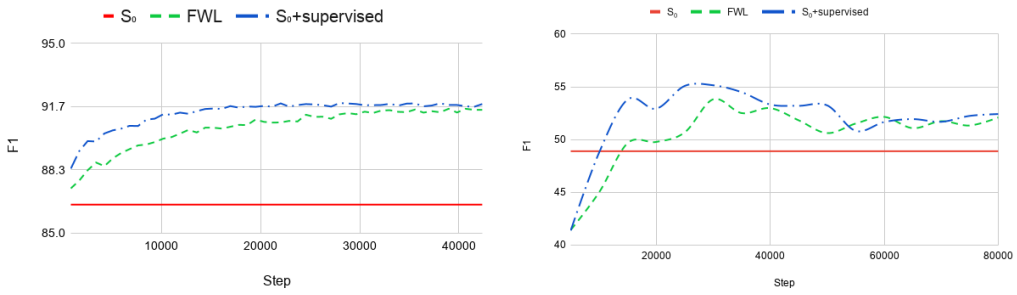
5.3a eta 5.3b Irudietako ikasketa kurbetatik ikus dezakegu nola portaera antzekoa den dokumentu sailkapen eta CQA atazetan. Bi kasuetan sistema gainbegiratuak FWLk baino azkarrago konbergitzen du. Hala ere, entrenamendua aurrera doan



(a) F1 puntuazioak *epoch* batean 3 lagin erabiliz.

(b) F1 puntuazioak 50 *epoch*etan 3 lagin erabiliz.

5.2 Irudia – Dokumentu sailkapenean hiperparametroen konbinazio desberdinak erabiliz garapen multzoan lortutako F1 puntuazioak erakusten dituzten beromapak. Antzeko errendimendua lortzen dugu hiperparametro bikote ezberdinekin, metodoaren sendotasuna erakutsiz.



(a) Dokumentu sailkapena

(b) CQA, testuingurua kontuan hartuz

5.3 Irudia – Dokumentu sailkapen eta CQA atazen ikasketa kurbak, non FWL ikasketa gainbegiratuarekin konparatzen den. FWLk S_0+ *gainbegiratura* konbergitzen du entrenamendu pausoak aurrera doazen ahala.

ahala, FWLn garapen azpi multzoko F1 balioak ere konbergitzen dute. Bereziki interesgarria da FWLk S_0 sistema hobetzen duen puntua. Hobekuntza hau lehen pausotan etortzen da bi kasuetan.

Mugak

Gure esperimentu guztietan erabiltzaile *feedback*a datu multzo gainbegiratu batekin simulatzen dugu, beraz, *feedback*a beti zehatza eta esplizitua izango da. Honetaz gain, ez dugu erabiltzaileak *feedback*a ematen ez duen kasua kontsideratzen eta errealitatean ezin dugu espero erabiltzaileak beti seinale hau ematea. Azkenik, gure metodoak laginketa erabiltzen duenez, erabiltzaile batek baina gehiagok galdera berdina egitea beharko genuke, hau muga argi bat izanik.

Laginketa eta ikasketa gainbegiratuaren arteko konparaketa

FWLn *epoch*ak ikasketa gainbegiratuan bezala tratatzen ditugunez, erantzun berri bat lagintzen dugu *epoch* bakoitzeko. Adibidez, dokumentu sailkapenaren kasuan guztira 150 lagin hartzen ditugu kontuan (50 *epoch* 3 laginekin) 219 klase bakarrik izanik. Argudiatu liteke modu sinple batean klase guztiak lagintzea klase zuzena edukitzearen baliokidea dela eta metodo hau FWL baino eraginkorragoa izango litzateke. Hala ere, S_0 sistema inguru errealista batean martxan jartzean laginketa sinple horrek oso probabilitate baxuko emaitzak laginduko lituzke, sistemaren erabilera baliogabetuz. Garrantzitsua da aipatzea galera funtzioaren gradientea kalkulatzen den kasu bakoitzean FWLk 3 laginen informazioa bakarrik duela, ikasketa gainbegiratuan, ordea, klase guztien inguruko informazioa dauka sistemak. Gainera, 3 lagin (*epoch* bat) nahikoak dira FWLk S_0 baina errendimendu hobea lortzeko (ikusi 5.2b Irudia).

6. KAPITULUA

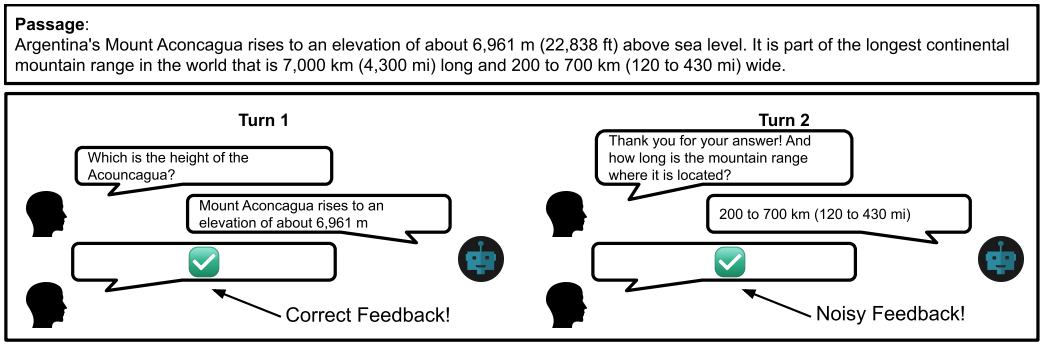
Erabiltzaile *feedback* bitar zaratatsua

Kapitulu honetan, aurrekoan bezala, aurretik entrenatua izan den galdera-erantzun sistema bat gizaki simulatuekin elkarrekintzan jarriko dugu. Atal honen helburua FWL algoritmoaren muga nagusi bat konpontzea da, erabiltzaile *feedback* seinaleak ekar dezakeen zarata modelatzea hain zuzen ere. Kapitulu honek aurreko kapituluak duen egitura bera jarraituko du: lehenik motibazioa eta ekarpenak, ondoren metodologia eta emaitzak eta azkenik metodoaren mugak azalduko ditugularik.

6.1 Motibazioa eta ekarpenak

Aurreko kapituluan erabiltzaile *feedback* bitarra erabiliz sistema bat hobetzeko FWL algoritmoa proposatu dugu. Algoritmo honek emaitza onak erakutsi ditu eszenatoki simulatu batean, baina erabilitako simulazioa errealitatetik urruti dago. Bertan, erabiltzaileen *feedback*a datu multzo estatiko batekin zuzenean simulatu da, hortaz, sistemak jasotako *feedback* guztia zuzena izango da, inongo zaratarik gabe. Metodo honek duen arazo nagusia da CQA eszenatoki erreal batean erabiltzaileek gehitzen duten zarata ez duela kontuan hartzen. Erabiltzaileak CQA sistema baten erantzuna jasotzean ez du beti argi izango erantzuna zuzena edo okerra den eta erabiltzailearen ezjakintasunaren ondorioz *feedback* bitarra zaratatsua izango da. Eszenatoki zaratatsu berri honen adibide bat 6.1 Irudian ikus daiteke.

Kapitulu honetan CQA sistemen gainean *feedback*a ematean erabiltzaileek du-



6.1 Irudia – Galdera-erantzunetan oinarritutako sistema baten adibide bat, non erabiltzaileak jasotzen duen erantzun bakoitzeko sistemari *feedback* bitar zaratatsua emango dion.

ten portaera aztertzen dugu. Erabiltzaile simulazio egoki bat egitea ezinbestekotzat jotzen dugu erabiltzaile *feedback* bitarra erabiliz ikasten duten algoritmoen errendimendua modu egokian ebaluatzeko. Hortaz, kapitulu honetako ekarpen nagusiak hurrengoak dira:

- Erabiltzaileen portaeraren azterketa CQA sistemei *feedback* bitarra ematean. Bertan, gizakiek gehitzen duten zarata kopurua estimatzen jarri dugu fokua.
- FWL algoritmoaren egokitzapena *feedback* zaratatsua kontuan har dezan. Algoritmoak ez badu zarata modelatzen, berdin sinetsiko ditu *feedback* seinale zuzen eta okerrak, errendimenduan ondorio negatiboak izanik.
- FWL algoritmo egokituaren analisia dokumentu sailkapen eta CQA atazetan.

6.2 Metodologia

Atal honetan 5. kapituluko metodologia berdina jarraituko dugu, hortaz, lehenik S_0 sistema gainbegiratu bat entrenatuko dugu ondoren erabiltzaileen *feedback* bitarra erabiliz hobetzen jarraitzeko. Aurreko kasuetan ez bezala, orain sistemak jasotzen duen *feedback* seinale bitarra zaratatsua izango da, beraz, gizakien elkarreraginetik gizakien galderak, sistemaren erantzunak eta gizakien *feedback* bitar zaratatsua jasotzen ditugu.

Galdera-erantzun sistemari dagokionez aurreko kasu guztietan bezala BERT sistemaren gainean bi sailkatzaile erabiltzen ditugu hasiera eta bukaera indizeak aurreikusteko. Kasu zaratatsuan sailkatzaile hauek entrenatzeko FWL algoritmoa eguneratzea proposatzen dugu *feedback*aren zuzentasuna adostasun mailaren bitartez modelatuz.

6.2.1 Ikasketa algoritmoa

Atal honetan FWL algoritmoaren aldaera bat aurkezten dugu adostasun maila erabiliz erabiltzaileengandik datorren zarata kontuan hartzeko diseinatua izan dena. Algoritmoak adostasun maila handia duten *feedback* seinaleak gehiago sinesten ditu adostasun maila txikia duten *feedback* seinaleak baino.

FWLren definizioan bezala, C klaseen gaineko eta x sarrera duen $p^*(y|x)$ distribuzioa definitzen dugu:

$$p^*(y|x) \propto exp(\beta f(y|x))$$

non $f(y|x) \in \{-1, 1\}$. *Feedback* determinista izan beharrean, oraingoan *feedback* estokastikoa kontsideratzen dugu zarata deskribatuko duen ϵ ausazko aldagai berri bat definituz. Ausazko aldagai berri honen ondorioz gure *feedback* funtzioa $f(y|x, \epsilon)$ izango da $f(y|x)$ izan beharrean. Kasu honetan ez dugu zarata ausazko aldagaiaren $p(\epsilon)$ distribuzioa ezagutzen baina asumitu dezakegu ϵ-en probabilitatea altua izango dela $f(y|x, \epsilon)$ zuzena izateko probabilitate handia denean.

Hau guztia zehaztu ostean, helburu distribuzioa honela definitu dezakegu:

$$p^*(y|x) = \int p^*(y|x, \epsilon)p(\epsilon)d\epsilon$$

FWL algoritmoan bezala, algoritmoaren helburua θ hiperparametroekiko p^*tik gure $q(y|x; \theta)$ sailkatzailerako KL dibergentzia minimizatzea da.

$$\text{KL}(p^*|q) = -\sum_y p^*(y|x) \log q(y|x; \theta) + \mathcal{H}(p^*). \tag{6.1}$$

FWLren deribazioan zehaztutako arrazoi berdinengatik, helburu funtzio honen minimizazio zehatza ezinezkoa izango da. Honen ordez, normalizatu gabeko

p^* kontsultatu dezakegu x sarrera eta y klase hautagai bat izanik. Hortaz, garrantzi bidezko laginketa autonormalizatua erabiltzen dugu hurrengo proposamen distribuzioaren bitartez:

$$\hat{q}(y|x) = \lambda q(y|x;\theta) + (1-\lambda)\mathcal{U}(y), \tag{6.2}$$

non $\mathcal{U}(y)$ $y$ren gaineko distribuzio uniforme bat den, q distribuzioa leuntzen lagunduko duena. Proposamen distribuzio honekin hurrengo helburu funtzioa deribatzen dugu 6.1 Ekuaziotik hasita:

$$\begin{aligned}
&\mathrm{KL}(p^*|q) - \underbrace{\mathcal{H}(p^*)}_{\text{const } \theta\text{rekiko}} = \\
&-\sum_y \hat{q}(y|x)\frac{\int p^*(y|x,\epsilon)p(\epsilon)d\epsilon}{\hat{q}(y|x)}\log q(y|x;\theta) > \\
&\quad -\sum_y \hat{q}(y|x)\frac{p^*(y|x,\epsilon)p(\epsilon)}{\hat{q}(y|x)}\log q(y|x;\theta) \\
&\qquad\qquad \because p(\epsilon) > 0 \text{ uneoro}
\end{aligned}$$

Honen gainean autonormalizatutako garrantzi bidezko laginketa aplikatuz, hurrengoa dugu:

$$\begin{aligned}
&\mathrm{KL}(p^*|q) \gtrapprox \\
&-\frac{1}{K}\sum_{k=1}^{K}\frac{\omega(y^k|x,\epsilon^k)}{\sum_{k=1}^{K}\omega(y^k|x,\epsilon^k)}\log q(y^k|x;\theta)
\end{aligned}$$

non K jasotako erabiltzaile *feedback* kopuru absolutua den. $\omega(y^k|x,\epsilon^k)$ garrantzi pisua honela konputatzen dugu:

$$\begin{aligned}
\log\omega(y^k|x,\epsilon^k) = \underbrace{\beta f(y^k|x,\epsilon^k)}_{=\text{feedback}} + \underbrace{\log p(\epsilon^k)}_{=\text{ zarata distribuzioa}} \\
- \underbrace{\log q(y^k|x)}_{=\text{ modeloaren sinesmena}},
\end{aligned}$$

Orain, $p(\epsilon^k)$ konputatzeko modu bat behar dugu. Azalpenaren hasieran zehaztu dugun bezala, $p(\epsilon^k)$ handia izan behar da k-garren *feedback*ak zarata gutxi duenean. Hortaz, $p(\epsilon^k)$ erabiltzaile *feedback*en adostasun mailarekiko modu proportzionalean definitzen dugu:

$$H(y) = \sum_{f \in \mathcal{F}} \frac{N_y^f + 1}{N_y + |\mathcal{F}|} \log \frac{N_y^f + 1}{N_y + |\mathcal{F}|}$$

non $\mathcal{F}$ *feedback* aukerak diren (e.g. $\{-1, 1\}$ *feedback* bitarrarentzat), N_y aurreikusitako klasearentzat jasotako *feedback* kopuru osoa den eta N_y^f aurreikusitako klasearentzat jasotako f motako *feedback* kopurua den. H termino hau entropia negatiboa bezala ikus daiteke. Hortaz, zarata probabilitatea honela definitu dezakegu:

$$p(\epsilon^k) \propto exp(\gamma H(y^k))$$

non γ hiperparametroak erabiltzaileen adostasunari ematen diogun garrantzia neurtuko duen. Hau guztia kontuan hartuta, $\omega(y^k|x, \epsilon^k)$ garrantzi pisu eguneratuak honela kalkulatuko ditugu:

$$\log \omega(y^k|x, \epsilon^k) = \underbrace{\beta f(y^k|x, \epsilon^k)}_{\text{=feedback}} + \gamma \underbrace{H(y^k)}_{\text{= adostasun maila}} - \underbrace{\log q(y^k|x)}_{\text{= modeloaren sinesmena}},$$

FWLren formulazio berri honek algoritmoa eguneratzen du zarata modelatu ahal izateko. Hitzez, gure algoritmoak erabiltzailearen *feedback*a laginak jasotako *feedback*aren adostasun mailarekiko modu proportzionalean erabiliko du. FWLren aldaera honi FWL-zaratatsua deitzen diogu.

6.2.2 Erabiltzaile simulazioa

Erabiltzaile *feedback*a informazio iturri merkea da sistema errealak martxan jartzerakoan, hau naturalki jaso badaiteke gizakien interakzioari esker. Laborategiko esperimentuak garatzeko, ordea, guztiz kontrakoa gertatzen da, esperimentu bakoitzean *feedback*a uneko sistemaren menpekoa baita. Honek esan nahi du sistema zehatz bat erabiliz *feedback* datu multzo estatiko bat biltzean posible dela hau

lagungarria ez izatea beste sistema batzuetarako, sistemak egiten dituzten akatsak desberdinak izan daitezkeelako.

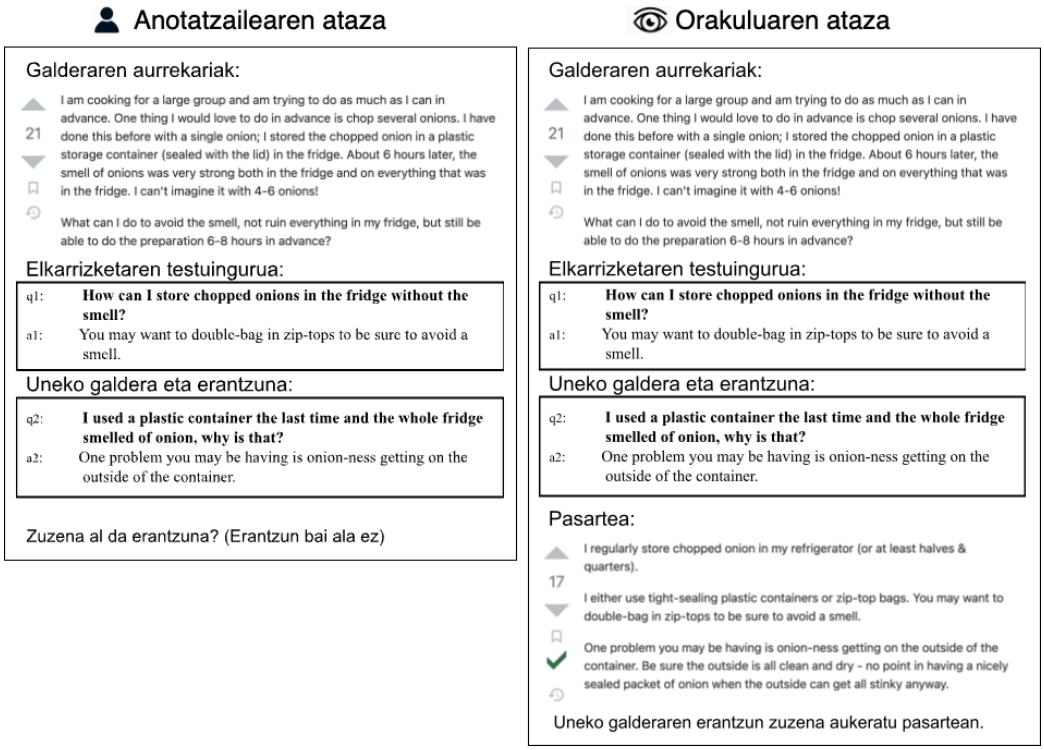
1 Algoritmoan datu multzo gainbegiratu batetik erabiltzaile *feedback*a simulatzeko oinarrizko metodoa azaldu dugu. Simulazio honek, ordea, arazo nagusi bat du, izan ere, ez du kontuan hartzen erabiltzaileetatik sistemak jasoko duen zarata. Zarata ez kontuan hartzeak FWL eta FWL-zaratatsua algoritmoen errendimendu erreala neurtzeko zailtasunak dakartza askotan errealitatean baino emaitza hobeak lortuz. Arazo hauei aurre egiteko, simulazio eszenatoki berri bat aztertzen dugu. Simulazio eszenatoki berri honi esker, *feedback* bitarra erabiltzen duten algoritmoen ebaluazio errealistago bat egin ahal dugu. Hau lortu ahal izateko lehen pausoa da erabiltzaileek CQA sistema bati *feedback*a ematean gehitzen duten zarata aztertzea.

Zarataren estimazioa

Erabiltzaileek CQA sistema bati *feedback* bitarra ematean duten portaera aztertzeko anotazio ataza berri bat definitu dugu. Ataza honetan, galderaren aurrekariak b, elkarrizketaren testuingurua eta uneko galdera eta erantzunak q_n, a_n emanda, anotatzaileak erabaki beharko du ea sistemak emandako erantzuna zuzena edo okerra den, hau da, *feedback* bitarra eman beharko du. Kasu honetan erantzunik gabeko (*I don't know* erantzuna dutenak) galderak baztertzen ditugu, erabiltzaileek ez baitaukate informazio nahikoa galdera bat p pasartean erantzungarria denetz jakiteko. Garrantzitsua da kontuan hartzea anotazio ataza honetan anotatzaileek erantzunak erauzteko erabiltzen den p pasartea ez dutela ikusten. Honela, ahalik eta gertuen gaude simulatu nahi dugun egoeratik, non erabiltzaile bat CQA sistema batekin elkarrekintzan dabilen eta honi *feedback* bitarra ematen dion. Anotazio ataza honetan anotatzaile artean lortutako adostasun maila % 80koa da.

Behin *feedback* anotazio guztiak ditugunean bertan dagoen zarata neurtu beharra daukagu. Honetarako aurreko pausoko anotazioak hartu eta orakulu anotatzaile bati pasatzen dizkiogu. Orakulu anotatzaileak erantzun zuzenak berriro anotatu behar ditu, $\hat{a_n}$ anotazio berria sortuz, galderaren aurrekariak b, elkarrizketaren testuingurua, uneko galdera eta erantzunak q_n, a_n eta p pasartea izanik. Anotatzaile honek orakulu izena jasoko du elkarrizketa informazio guztia atzitu dezakeelako. 6.2 Irudian ikus daitekeen bezala, orakulu anotatzaileak aurretik azaldutako informazio guztia atzigarri izanik erantzun zuzena anotatuko du pasartean.

Prozesu hau formalizatzeko, O orakulu eta U erabiltzaile ausazko aldagaiak definitzen ditugu. Honetaz gain, $\omega = \{\textit{zuzen}, \textit{oker}\}$ lagin espazioa zehaztuko dugu, non



6.2 Irudia – *Feedback* bitar zaratatsua estimatzeko atazaren adibide bat. Bertan anotatzaileak *feedback* bitarra eman beharko du. Orakuluak, ordea, erantzuna berriro anotatu beharko du pasarte originalean azpi pasarte bat aukeratuz. Orakuluak erantzun zuzena aukeratzeko beharrezkoa den informazio guztia atzigarri dauka.

$$O(\omega) = \begin{cases} 1, & \text{orakuluak zuzena dela dionean} \\ 0, & \text{orakuluak okerra dela dionean} \end{cases}$$

$$U(\omega) = \begin{cases} 1, & \text{erabiltzaileak zuzena dela dionean} \\ 0, & \text{erabiltzaileak okerra dela dionean} \end{cases}$$

Formalizazio hau kontuan izanik, anotazio prozesu honen bidez $P(\neg U|O)$ eta $P(U|\neg O)$ estimatu nahi ditugu, hauek baitira kasu zaratatsuak. Estimazioak aurrera eramateko erabiltzaileek anotatutako a_n erantzuna eta orakuluak sortutako

$\hat{a_n}$ erantzunak konparatzen ditugu. Bi erantzunen arteko F1 neurria $0,2$ baina txikiagoa denean, $a_n$ri orakuluak *feedback* negatiboa ematen diola kontsideratzen dugu. F1 neurria $0,8$ baina handiagoa den kasuetan, ordea, orakuluak *feedback* positiboa ematen duela zehaztuko dugu. Kasu honetan $0,2$ eta $0,8$ko balioak hartzen ditugu anotatutako kasu guztien % 93a osatzen dutelako. Behin orakuluaren eta erabiltzaileen *feedback* seinaleak ditugunean $P(\neg U|O)$ eta $P(U|\neg O)$ zaratak estimatu ditzakegu hurrengo ekuazioen bitartez:

$$P(\neg U|O) = \frac{\{\text{orakuluarentzat zuzena}\} \cap \{\text{erabiltzailearentzat okerra}\}}{\{\text{orakuluarentzat zuzena}\}}$$
$$P(U|\neg O) = \frac{\{\text{orakuluarentzat okerra}\} \cap \{\text{erabiltzailearentzat zuzena}\}}{\{\text{orakuluarentzat okerra}\}}$$

Estimazio hauetatik lortzen ditugun balioak $P(\neg U|O) = 0,21$ eta $P(U|\neg O) = 0,25$ dira. Balio hauetatik ondorioztatu dugu erabiltzaile batek erantzun zuzen bati *feedback*a ematen dion bakoitzean kasuen $\sim$ %21a zaratatsua izango dela. Modu antzekoan, erantzun oker bati *feedback*a ematean kasuen $\sim$ %25a zaratatsua izango da.

Simulazio algoritmoa

Feedback bitarra ematean erabiltzaileek gehitzen duten zarata aztertu ostean, prozesu hau simulatzeko garatu dugun algoritmo eguneratua definituko dugu. Honetarako, erabiltzaileek gehitzen duten zarata *Bernoulli* distribuzio baten bitartez modelatzen dugu, aurreko atalean estimatutako $P(\neg U|O) = 0,21$ eta $P(U|\neg O) = 0,25$ balioak erabiliz.

2 Algoritmoaren lehen pausoan q_i galderarentzat lagindutako $a_{i,j}$ erantzunaren eta datu multzoko $\tilde{a_i}$ erantzunaren arteko $F1$a kalkulatzen dugu. Ondoren, $F1$a th_{pos} balioaren berdina edo handiagoa bada, *Bernoulli* distribuzio batetik laginduko dugu p_{pos} arrakasta probabilitatearekin. Gure esperimentuetan $p_{pos} = P(\neg U|O)$ bezala definitu dugu. *Bernoulli* lagina arrakastatsua denean, erantzun zuzen bati *feedback* negatiboa emanez zarata gehituko dugu. $F1$ neurria th_{neg} balioaren berdina edo txikiagoa denean, antzeko prozesu bat jarraitzen dugu baina $p_{neg} = P(U|\neg O)$ izanik. Kasu guztietan $th_{pos} = 1,0$ eta $th_{neg} = 0,0$ izango dira, beraz, erantzuna positiboa dela esango dugu bakarrik datu multzokoaren berdina den kasuetan.

Algorithm 2: *Feedback* zaratsua datu multzo estatiko batekin simulatzeko algoritmoa. Bertan, *Bernoulli* distribuzio bat erabiliz, jasotako *feedback* bitar seinalea irauli egiten da.

Input: $\hat{a}_i$, $a_{i,j}$, th_{pos}, th_{neg}, p_{pos}, p_{neg}
Output: F_i (i galderarentzat jasotako *feedback*ak)
$F1 \rightarrow F1(a_{i,j}, \hat{a}_i)$;
if $F1 \geq th_{pos}$ **then**
    if $Bern(p_{pos})$ **then**
        FeedbackPositiboaEman(F_i);
    end
    else
        FeedbackNegatiboaEman(F_i);
else if $F1_k \leq th_{neg}$ **then**
    if $Bern(p_{neg})$ **then**
        FeedbackPositiboaEman(F_i);
    end
    else
        FeedbackNegatiboaEman(F_i);

6.3 Esperimentuak

Azpiatal honetan FWL algoritmoaren aldaera berria ebaluatzeko esperimentuak aurkezten ditugu. Esperimentu bakoitzean honen helburua, atazaren deskribapena, erabilitako sistemak eta emaitzak azalduko ditugu aurreko kasuetan bezala.

5.3 atalean bezala, kasu honetan ere, S_0 sistema gainbegiratu batetik hasten gara. Ondoren sistema honi esker, errealitatean sistema bat gizakiekin martxan jarri eta geroko fasea simulatzen dugu. Aurreko atalarekin alderatuz, desberdintasun nagusia erabiltzailea simulatzeko erabiltzen dugun algoritmoa izango da, izan ere, 6.2.2 azpiatalean definitutako simulazio algoritmo berria erabiliko dugu. Algoritmo honek *feedback* perfektua eman beharrean erabiltzaileek gehitzen duten zarata simulatuko du, eszenatoki errealistago bat sortuz. Entrenamendu eta simulaziorako erabiltzen ditugun multzoak aurreko ataleko berdinak izango dira. Beraz, kasu honetan esperimentuetarako zehazten ditugun sistemak aurreko atalekoen oso antzekoak izango dira, S_0, S_0 + *gainbegiratua* eta guztiz gainbegiratutako sistema berdin mantentzen direlarik. Aldaera berriak hurrengoak dira:

- S_0 + *FWL*: jatorrizko S_0 sistema FWL algoritmoarekin birdoitzen dugu,

baina kasu honetan simulazio azpimultzoko adibideak simulazio algoritmo berriarekin erabiltzen ditugu *feedback* bitar zaratatsua lortuz.

- S_0 + *FWL-zaratsua*: jatorrizko S_0 sistema FWL-zaratatsua algoritmoarekin birdoitzen dugu, simulazio azpimultzoko adibideak simulazio zaratatsuaren gainean aplikatuz.

Hiperparametroen doikuntzarako dokumentu sailkapena erabiltzen dugu berriro ere ataza lagungarri bezala. Kasu honetan, ordea, optimizatu beharreko hiperparametroak hiru dira: β, λ eta γ. Kontuan izan, γ hiperparametroak adostasun mailari emandako garrantzia kontrolatuko duela.

6.3.1 Dokumentu sailkapena

Helburua

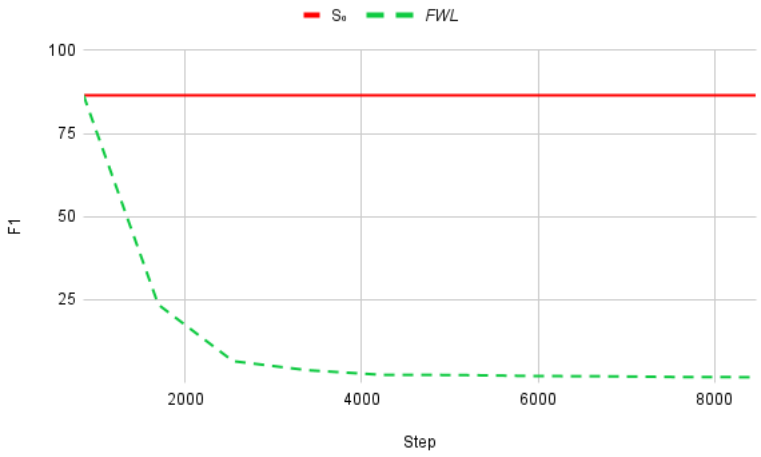
Dokumentu sailkapen atazan egindako esperimentuen helburua FWL-zaratatsua algoritmoaren hiperparametroen doikuntza egitea da. Honekin batera, proposatutako aldaeraren hobekuntzak erakusteko, FWL eta FWL-zaratsuaren arteko konparaketa egin nahi da. Algoritmoaren aldaerak zarata modu egokian modelatzea lortzen badu, FWL baina sendoagoa dela ikusi beharko litzateke garatu ditugun esperimentuetan.

Atazaren deskribapena

Kasu honetan ere *DBPedia Classes* datu multzoa erabiltzen dugu non x dokumentu bat izanik, sistemak dokumentuaren y klase zuzena aurreikusi behar duen. Erabilitako S_0 sistema berriro ere jatorrizko datuen % 10arekin entrenatu dugu. Ondoren, simulazio garaian datu multzoaren % 90a erabiltzen jarraitzen dugu, baina simulazio algoritmoa 2 Algoritmoan aurkezten duguna izango da. Hortaz, kasu honetan sistemak simulazio garaian jasoko duen *feedback*a zaratatsua izango da.

Erabilitako sistemak

Esperimentu hau aurrera eramateko aurreko kapituluko esperimentuetako *MLP* bera erabili dugu. Hortaz, eredu honek geruza ezkutu bakar bat izango du eta sarrerako errepresentazioa dokumentuko hitz-bektoreen batezbestekoa izango da.



6.3 Irudia – FWL algoritmoaren entrenamendu dinamika *feedback* zaratatsuaren kasuan. Bertan argi ikus daiteke degenerazioaren kasua, non geroz eta *feedback* gehiago jaso sistemak errendimendu txarragoa duen.

Emaitzak

Entropia gurutzatua erabiliz entrenatutako S_0 sistemaren emaitzak aurreko kapituluan aurkeztutakoen berdinak dira, esperimentu honetan ez baitago inongo desberdintasunik. Gauza bera gertatzen da S_0 + *gainbegiratua* eta *guztiz gainbegiratua* kasuetan. S_0 + *FWL* sistemarentzat $\lambda = 0,97$ eta $\beta = 76$ hiperparametroak erabiltzen ditugu. S_0 + *FWL-zaratatsua* algoritmoarentzat, ordea, γ hiperparametroaren esplorazio bat egiten dugu optimizazio bayestarraren bitartez, beste bi hiperparametroak $\lambda = 0,97$ eta $\beta = 76$ bezala mantentzen ditugularik. Esplorazio honetan lorturiko balio onena $\gamma = 256$ da. Beste entrenamenduko parametroak berdin mantentzen ditugu, klase kopurua 3 izanik eta *epoch* kopurua 50ean zehaztuz, beraz, *epoch* bat 3 klase lagintzearen baliokidea izango da.

6.1 Taulan ikus daiteke nola FWL algoritmoa bere baitan ez den gai simulatutako zarata maneiatzeko, sistemaren degenerazioa bultzatuz. Sistemaren degenerazioa da martxan jarri eta gero sistema berria jatorrizko S_0 sistema baina okerragoa denean. Honekin batera, garrantzitsua da aipatzea nola behin sistema degeneratzen hasten denean hau berreskuratzea ezinezkoa izaten den, laginen kalitatea okertzen doan ahala *feedback*aren erabilgarritasuna ere okertzen baitoa. Ikasketa dinamikak azaltzen dituen 6.3 Irudian ikus daiteke degenerazio fenomenoaren adibide argi bat.

Sistema	F1
S_0	86,51
S_0 + *FWL*	1,65 (-84,86)
S_0 + *FWL-zaratatsua*	90,22 (+3.71)
S_0 + *gainbegiratua*	91,89 (+5,3)
Guztiz gainbegiratua	92,04 (+5,5)

6.1 Taula – Dokumentu sailkapeneko emaitzak. FWL-zaratatsua algoritmoak S_0 hobetzen jarraitzen du *feedback* bitarra bakarrik erabiliz, hau zaratsua den kasuetan ere.

Bestalde, S_0 + *FWL-zaratatsua* algoritmoaren emaitzak aztertzen baditugu, ikus daiteke nola proposatutako algoritmo berriak emaitza onak lortzen dituen eta gai den S_0ren emaitzak hobetzeko. Emaitza hauek erakusten dute, nola proposatutako algoritmoaren aldaera sendoagoa den erabiltzaile zarataren aurrean. Hala ere, garrantzitsua da nabarmentzea nola S_0rekiko hobekuntza txikiagoa den testuinguru zaratatsuan. 5.3.1 atalean, *feedback*a zaratsua ez denean, FWLrekin S_0rekiko lortu dugun hobekuntza 5,0 puntuko F1ekoa da, kasu honetan, ordea, 3,71ko hobekuntza lortzen dugu.

6.3.2 CQA ataza

Helburua

Esperimentu honen helburua da FWL-zaratatsua algoritmoaren portaera aztertzea CQA atazaren gainean erabiltzaile zarata simulatua gehitzen dugunean. Dokumentu sailkapenean lortutako emaitza onak positiboak diren arren, errealitatetik gertuago dagoen ataza konplexuago batean ebaluatzea ezinbestekoa da.

Atazaren deskribapena

Esperimentu hauetarako QuAC datu multzoa erabiltzen dugu, aurreko kasuetan bezala % 10 entrenamendurako erabiliz eta % 90 simulaziorako erabiliz. Desberdintasun nagusia simulazio garaian gehitzen dugun zaratan datza, beraz, martxan jarri eta geroko FWLren bi aldaerek *feedback* zaratatsua jasoko dute kasu honetan. Algoritmoen hiperparametroei dagokienez, 6.3.1 atalean lortutako konbinazio onena erabiltzen dugu: $\lambda = 0,97$, $\beta = 76$ eta $\gamma = 256$. Lagin kopuruari dagokionez, 50en mantentzen dugu aurreko CQA esperimentuetan bezala.

Sistemak	Testuinguruarekin
S_0	49,03
S_0 + *FWL*	2,8 (- 46,23)
S_0 + *FWL-zaratatsua*	3,9 (- 45,13)
S_0 + *gainbegiratua*	55,10 (+6,1)
Gainbegiratu osoa	55,40 (+6,5)

6.2 Taula – Domeinu barneko esperimentuen emaitzak *feedback*a zaratatsua den kasuan. Bertan, QuAC datu multzoa erabiltzen dugu entrenamendu eta simulaziorako, bigarren fase honetan zarata gehitzen dugularik. F1 balioak QuACeko garapen azpi multzoan kalkulatzen ditugu. Kasu honetan, *feedback* zaratatsua bakarrik erabiltzen duten sistema guztiak degeneratu egiten dute.

Erabilitako sistemak

Berriro ere ataza hau ebazteko BERTen galdera-erantzun atazarako aldaera erabiltzen dugu. Zarataren eragina aztertzeko testuingurua kontuan hartzen duen aldaera bakarrik erabiliko dugu, beraz, uneko galderari testuinguruko galdera eta erantzunak kateatuko dizkiogu. Esperimentu hauetan eredu hau bakarrik erabiltzen dugu 5. Kapituluan emaitza onenak lortzen baititu.

Emaitzak

6.2 taulan ikus daitekeen bezala, FWL algoritmoaren bi aldaerek degeneratu egiten dute jasotako *feedback*ak zarata duen kasuetan. Emaitza hauek dokumentu sailkapenekoetatik oso desberdinak dira eta erakusten dute nola ataza konplexuago batean *feedback* zaratatsua maneiatzeko proposatutako FWL-zaratatsua algoritmo berria ez den nahikoa. Emaitza negatibo hauek proposatutako metodoaren sendotasun falta erakusten dute, bereziki ataza konplexu eta errealistetan aplikatzerakoan. Esperimentu hauek ebaluazioaren garrantzia azpimarratzen dute, dokumentu sailkapenaren ataza sinpleak proposatutako algoritmo berriaren errendimendua gain baloratzea eragin baitute.

6.4 Ondorioak

Garatutako dokumentu sailkapen esperimentuetan erakutsi dugu nola FWL+zarata algoritmoa gai den erabiltzaileen simulaziotik datorren zarata modelatzeko eta S_0

sistema bat hobetzeko jasotako *feedback*ak zarata kopuru errealista bat duenean. Hemendik ondoriozta daiteke FWL+zarata algoritmoa behar bezain sendoa dela sistema errealetan inplementatzeko. CQA atazan egindako esperimentuetan, ordea, argi ikusten da nola algoritmo berria ez den gai zarata honetatik informazio egokia erauzteko. Metodoaren errendimenduaren erorketa CQA atazaren konplexutasunaren ondorioz gertatzen da, izan ere, bigarren ataza honetan S_0 sistemaren jatorrizko kalitatea ($F1 = 46,01$) dokumentu sailkapenean ($F1 = 86,51$) baina askoz baxuagoa da. Hortaz, sistemaren jatorrizko kalitatea baxua den kasuetan zaratak ondorio larriagoak izan ditu. Honetaz gain, erabiltzailea simulatzeko diseinatu dugun algoritmoan zarata ausaz gehitzen dugu nahiz eta zarata portzentaia erabiltzaileetatik estimatzen dugun. Ausazkotasun honek zarataren modelatzea asko zailtzen du.

Esperimentu hauetan ikusi dugu nola erabiltzaile *feedback*etik ikastea oso ataza konplexua den, hau modu egokian aztertzeko zenbait baldintza nabarituz. Lehenik, ezinbestekoa da sistema batek errealitatean topatuko dituen zailtasunak islatzen dituen datu multzo bat izatea, erabilitako dokumentu sailkapen atazan gertatzen ez den bezala. Honetaz gain, garrantzi handikoa da erabiltzaile simulazioa egiteko algoritmoak ausazkoak ez izatea, sistemek mundu errealean ez baitute ausazko *feedback*ik jasoko. Adibidez, gizakiekin martxan dagoen sistema batek zarata handiagoa jasoko du galdera zailen *feedback*ean galdera errazenean baino. Gure simulazio algoritmoan, ordea, kasu guztietan zarata kopurua berdina izango da, honen modelatzea asko zailduz.

7. CHAPTER

Conclusion and future work

In this thesis, we have defined the first steps towards using binary user feedback for improving conversational QA systems. In addition, we have shown that current technology has difficulties when handling the noisy feedback signal. Ultimately, the dataset developed during this thesis, along with other contemporaneous ones, has played a central role in the improvement on the quality of CQA systems. More concretely, the main **contributions** of this thesis are the following ones:

- We have created a **new conversational QA dataset** known as DoQA that enables the access to domain specific FAQs via conversations. This dataset contains 2,437 information-seeking dialogues on three different domains that are the cooking, travel and movies domains for a total of 10,917 questions. The dialogues in this dataset are created by crowd workers that play two different roles: (1) the user that asks questions about a certain topic posted in Stack Exchange; (2) the domain expert that replies to the questions by selecting a span of text from the original reply in the Stack Exchange post. Compared to previous CQA datasets, DoQA responds to real information needs, is multi-domain and more natural and coherent. Moreover, we present an ODQA setting for the dataset that was not part of the previously developed CoQA and QuAC datasets. Our efforts for creating a more realistic CQA dataset and publicly releasing it has enabled further research in CQA bringing the improvement of CQA models and motivaintg the creation of more challenging datasets. This is related with the **RL1** presented in 1.2.

- We have conducted a **thorough analysis** of the capabilities and limitations of the previous state-of-the-art CQA models on the DoQA dataset. We paid especial attention to the out-of-domain case as this is a very common issue that deployed NLP systems face. Here, we showed how BERT models performance dropped slightly (~ 3 F1) in the out-of-domain setting. However, due to the complexity of the DoQA dataset, we showed how the baseline PLMs are still far from humans in this task. This big gap motivated future research on CQA (Chen *et al.* 2021; Gekhman *et al.* 2022). For instance, Gekhman *et al.* (2022) introduced a new prompt-based history modeling approach called MarCQAp that was able to outperform all previous systems in the DoQA dataset. This is related with the **RL2** presented in 1.2.

- We have **proposed a new algorithm** (FWL) that allows a supervised classifier to effectively adapt itself after deployment just using binary user feedback. Moreover, we designed a simulation algorithm for human binary feedback starting from a static dataset that enables controlled synthetic experiments. We showed how our proposed algorithm is effective in the synthetic environment we designed on top of two CQA datasets. It is of especial interest the case of out-of-domain deployment, where the classifier is able to improve in a new domain via partial feedback coming from the interaction with simulated perfect users. We also presented positive results with two different architectures, including a multi-layer feed forward network and a PLM showing the robustness of the method in this setting. On top of that we performed an in-depth analysis of the limitations of the method in terms of sampling efficiency and the lack of noise in the feedback. We highlight that addressing this points is essential for real deployment of our algorithm. This is related with the **RL2** and **RL3** presented in 1.2.

- We have **designed an alternative simulation and learning from feedback algorithm** motivated by the limitations of the initial approach. In order to improve the simulation, we have performed a user study for modeling the noise that users incorporate when giving feedback on the CQA task. Once having an estimation of this noise, we have updated the simulation algorithm for incorporating noise when giving feedback. In order to handle the new noise coming from the feedback labels we have updated the initial FWL algorithm. The updated version of FWL now adapts the supervised classifier in a proportional way to the user agreement on the feedback labels. This updated algorithm has shown to be effective in the easy task of

document classification, but not when applying it to the more complex task of CQA. These negative results show how handling user noise is still a challenging and open research direction. This is related with the **RL2** and **RL3** presented in 1.2.

In terms of **publications**, this thesis contains 2 papers published in international conferences: ACL and COLING. The paper presented at COLING 2020 received a **best paper award nomination**. In addition, we published 8 other peer reviewed papers during this PhD (2 LREC, 1 EMNLP, 2 Ikergazte, 1 SemEval, 2 ACL Workshops), including a **best paper award nomination** at EMNLP 2020 and a most relevant research for the development of the Basque Country **award at IkerGazte** 2021. Two other papers are currently under review for ICML.

The COVID-19 pandemic has presented challenges for the scientific community, and we have used techniques from our thesis to help address some of these challenges. Our research team participated in two challenges related to COVID-19: the CORD-19 challenge on Kaggle and the EPIC QA challenge at TAC 2020. The CORD-19 (COVID-19 Open Research Dataset Challenge) competition was jointly organized by multiple organizations, including the Allen Institute for AI, Chan Zuckerberg Initiative, Georgetown University, Microsoft Research, National Institutes of Health, and The White House Office of Science and Technology Policy. This initiative made over 50,000 scientific articles related to COVID-19, SARS-CoV-2, and other coronaviruses available to the global research community. In this CORD-19 challenge, we developed text and data mining tools to help the medical community answer important scientific questions, and our team **won one of the nine tasks** [1]. On the other hard, the EPIC-QA challenge was hosted in the Thirteenth Text Analysis Workshop (TAC 2020). The aim of the EPIC-QA track was to assess the proficiency of systems in delivering comprehensive responses to inquiries concerning the disease COVID-19, its underlying virus SARS-CoV-2, other relevant coronaviruses, and the suggested measures to address the pandemic. In this QA challenge, we fine-tuned a scientific model on QA data and ranked 3rd in the competition. This model has been downloaded over 35,483 times from the Huggingface (Wolf *et al.* 2020) Hub and has the potential to advance research in the field of domain-specific QA.

Furthermore, during the development of the thesis, we have also contributed to the generation of **valuable resources** for the Basque language. Specifically, we created BERTeus (Agerri *et al.* 2020), a BERT base model designed exclusively

[1] `https://www.kaggle.com/code/aotegi/neural-question-answering-for-cord19-task8/notebook`

for the Basque language, and IxamBERT (Otegi *et al.* 2020), a Multilingual BERT model that integrates Basque, Spanish, and English. We showed that our pre-trained models outperformed previously available multilingual PLMs containing the Basque language in their pre-training. The impact of our research can already be seen in the number of downloads that these models have received from the Huggingface model hub. BERTeus has been downloaded 69,049 times, while mBERTeus has been downloaded 39,645 times, demonstrating the potential of our work for advancing language technology in underrepresented languages [2] [3]. Moreover, we also developed a small scale CQA dataset for the Basque language that has been used for the evaluation of natural language understanding (NLU) in Basque (Urbizu *et al.* 2022; Artetxe *et al.* 2022).

Future Work

Going back to the work presented in the manuscript, it is remarkable how the usage of human feedback has gain a great importance in the last year. For instance, ChatGPT, that is the fastest growing machine learning model ever deployed reaching 1,000,000 users in 5 days, was fine tuned using reinforcement learning from human feedback (Christiano *et al.* 2017). We expect that the usage of human feedback as a signal for improving deployed ML systems will become one of the main research areas and will make significant progress in the near future. More concretely, the main research lines that we would like to explore are the following ones:

- Our study has demonstrated the difficulties involved in modeling user feedback noise, and the significant harm it can inflict on a system, potentially causing it to become entirely dysfunctional. Despite these challenges, we consider noisy human feedback to be an accessible and inexpensive signal for deployed systems, as it requires minimal effort from users. While current techniques involve using instructed annotators to gather human feedback datasets, this approach is costly, limits scalability, and is restricted to the preferences of the annotators. Drawing inspiration from the success of deployed assistants like ChatGPT, we believe that exploring effective ways to handle noisy user feedback and utilize these signals is a crucial area for research.

[2] (Last accessed on 04/07/2023)

[3] All the developed models can be accessed in `https://huggingface.co/models`

- Recent research has demonstrated the effectiveness of learning from natural language feedback as a potential alternative to the widely used RLHF (Reward Learning from Human Feedback) approach (Scheurer *et al.* 2023; Chen *et al.* 2023). RLHF trains a reward model by having humans compare system outputs and selecting the better one. However, we believe that natural language feedback provides more information than human comparisons. Nonetheless, there has been no formal analysis conducted to compare the advantages and disadvantages of the two methods. Therefore, we argue that a formal analysis of these two approaches is crucial and an interesting direction for future research.

- As previously mentioned, studies have revealed that using natural language feedback can enhance machine learning (ML) systems significantly. However, requesting such feedback from deployed system users is not realistic. Additionally, obtaining natural language feedback from trained annotators is an expensive affair. Motivated by this, we would like to analyze and use language feedback generated by models in the future. While some preliminary work in this direction exists (Bai *et al.* 2022; Madaan *et al.* 2023), there is a lack of analysis regarding the constraints of this approach, particularly when compared to feedback provided by humans.

Bibliography

Aceta C., Fernández I., and Soroa A. Todo: A core ontology for task-oriented dialogue systems in industry 4.0. *Further with Knowledge Graphs*, 1–15. IOS Press, 2021.

Adlakha V., Dhuliawala S., Suleman K., de Vries H., and Reddy S. Topiocqa: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483, 2022.

Agerri R., San Vicente I., Campos J.A., Barrena A., Saralegi X., Soroa A., and Agirre E. Give your text representation models some love: the case for Basque. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4781–4788, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.588.

Agirre E., Jonsson A., and Larcher A. Framing Lifelong Learning as Autonomous Deployment: Tune Once Live Forever. *International Workshop on Spoken Dialogue Systems Technology*, 2019.

Anantha R., Vakulenko S., Tu Z., Longpre S., Pulman S., and Chappidi S. Open-domain question answering goes conversational via question rewriting. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 520–534, 2021.

Artetxe M., Aldabe I., Agerri R., Perez-de Viñaspre O., and Soroa A. Does corpus quality really matter for low-resource languages? *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7383–7390,

Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.499.

Askell A., Bai Y., Chen A., Drain D., Ganguli D., Henighan T., Jones A., Joseph N., Mann B., DasSarma N., *et al.*. A General Language Assistant as a Laboratory for Alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Bai Y., Kadavath S., Kundu S., Askell A., Kernion J., Jones A., Chen A., Goldie A., Mirhoseini A., McKinnon C., *et al.*. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Black S., Biderman S., Hallahan E., Anthony Q., Gao L., Golding L., He H., Leahy C., McDonell K., Phang J., *et al.*. Gpt-neox-20b: An open-source autoregressive language model. *Proceedings of BigScience Episode\# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, 95–136, 2022.

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., *et al.*. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. URL https://arxiv.org/pdf/2005.14165.pdf.

Bubeck S., Chandrasekaran V., Eldan R., Gehrke J., Horvitz E., Kamar E., Lee P., Lee Y.T., Li Y., Lundberg S., *et al.*. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Campos J.A., Cho K., Otegi A., Soroa A., Agirre E., and Azkune G. Improving conversational question answering systems after deployment using feedback-weighted learning. *Proceedings of the 28th International Conference on Computational Linguistics*, 2561–2571, 2020.

Castelli V., Chakravarti R., Dana S., Ferritto A., Florian R., Franz M., Garg D., Khandelwal D., McCarley S., McCawley M., Nasr M., Pan L., Pendus C., Pitrelli J., Pujar S., Roukos S., Sakrajda A., Sil A., Uceda-Sosa R., Ward T., and Zhang R. The TechQA Dataset, 2019.

Charniak E., Altun Y., de Salvo Braz R., Garrett B., Kosmala M., Moscovich T., Pang L., Pyo C., Sun Y., Wy W., *et al.*. Reading comprehension programs in a statistical-language-processing class. *ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, 2000.

Chen A., Scheurer J., Korbak T., Campos J.A., Chan J.S., Bowman S.R., Cho K., and Perez E. Improving code generation by training with natural language feedback. *arXiv preprint arXiv:2303.16749*, 2023.

Chen D. *Neural reading comprehension and beyond*. Stanford University, 2018.

Chen Y., Wu L., and Zaki M.J. Graphflow: exploiting conversation flow with graph neural networks for conversational machine comprehension. *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 1230–1236, 2021.

Choi E., He H., Iyyer M., Yatskar M., Yih W.t., Choi Y., Liang P., and Zettlemoyer L. QuAC: Question Answering in Context. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2174–2184, 2018.

Christiano P.F., Leike J., Brown T., Martic M., Legg S., and Amodei D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. URL `https://arxiv.org/pdf/1706.03741.pdf`.

Clark K., Luong M.T., Le Q.V., and Manning C.D. Electra: Pre-training text encoders as discriminators rather than generators. *International Conference on Learning Representations*, 2020.

Dai Z., Zhao V.Y., Ma J., Luan Y., Ni J., Lu J., Bakalov A., Guu K., Hall K.B., and Chang M.W. Promptagator: Few-shot dense retrieval from 8 examples, 2023. URL `https://arxiv.org/abs/2209.11755`.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., and Harshman R. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Deriu J., Rodrigo A., Otegi A., Echegoyen G., Rosset S., Agirre E., and Cieliebak M. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810, 2021.

Devlin J., Chang M.W., Lee K., and Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186, 2019.

Di Langosco L.L., Koch J., Sharkey L.D., Pfau J., and Krueger D. Goal misgeneralization in deep reinforcement learning. *International Conference on Machine Learning*, 12004–12019. PMLR, 2022.

Dunn M., Sagun L., Higgins M., Guney V.U., Cirik V., and Cho K. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. *arXiv preprint arXiv:1704.05179*, 2017.

Elgohary A., Peskov D., and Boyd-Graber J. Can you unpack that? learning to rewrite questions-in-context. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5918–5924, 2019.

Feng S., Patel S.S., Wan H., and Joshi S. Multidoc2dial: Modeling dialogues grounded in multiple documents. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6162–6176, 2021.

Feng S., Wan H., Gunasekara C., Patel S., Joshi S., and Lastras L. doc2dial: A goal-oriented document-grounded dialogue dataset. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8118–8128, 2020.

Furnas G.W., Landauer T.K., Gomez L.M., and Dumais S.T. The vocabulary problem in human-system communication. *Communications of the ACM*, 30 (11):964–971, 1987.

Gao G., Choi E., and Artzi Y. Simulating bandit learning from user feedback for extractive question answering. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5167–5179, 2022.

Gao L., Biderman S., Black S., Golding L., Hoppe T., Foster C., Phang J., He H., Thite A., Nabeshima N., *et al.*. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gao L. and Callan J. Unsupervised corpus aware language model pre-training for dense passage retrieval. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2843–2853, 2022.

Gehman S., Gururangan S., Sap M., Choi Y., and Smith N.A. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369, 2020.

Gekhman Z., Oved N., Keller O., Szpektor I., and Reichart R. On the robustness of dialogue history representation in conversational question answering: A comprehensive study and a new prompt-based method. *arXiv preprint arXiv:2206.14796*, 2022.

Géry M. and Largeron C. Bm25t: a bm25 extension for focused information retrieval. *Knowledge and information systems*, 32:217–241, 2012.

Guo M., Zhang M., Reddy S., and Alikhani M. Abg-coqa: Clarifying ambiguity in conversational question answering. *3rd Conference on Automated Knowledge Base Construction*, 2021.

Guu K., Lee K., Tung Z., Pasupat P., and Chang M.W. Realm: retrieval-augmented language model pre-training. *Proceedings of the 37th International Conference on Machine Learning*, 3929–3938, 2020.

Hancock B., Bordes A., Mazare P.E., and Weston J. Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415*, 2019.

He H., Balakrishnan A., Eric M., and Liang P. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1766–1776, 2017.

Hirschman L., Light M., Breck E., and Burger J.D. Deep read: A reading comprehension system. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, 325–332, 1999.

Hoi S.C., Sahoo D., Lu J., and Zhao P. Online learning: A Comprehensive Survey. *arXiv preprint arXiv:1802.02871*, 2018.

Huang H.Y., Choi E., and Yih W.t. Flowqa: Grasping flow in history for conversational machine comprehension. *International Conference on Learning Representations*, 2019.

Huang H.Y., Zhu C., Shen Y., and Chen W. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. *International Conference on Learning Representations*, 2018.

Humeau S., Shuster K., Lachaux M.A., and Weston J. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. *International Conference on Learning Representations*, 2020.

Izacard G., Caron M., Hosseini L., Riedel S., Bojanowski P., Joulin A., and Grave E. Towards unsupervised dense information retrieval with contrastive learning. *CoRR*, abs/2112.09118, 2021. URL `https://arxiv.org/abs/2112.09118`.

Jurafsky D. and Martin J.H. *Speech and language processing*, 3 lib. Pearson London, 2014.

Karpukhin V., Oguz B., Min S., Wu L., Edunov S., Chen D., and Yih W. Dense passage retrieval for open-domain question answering. *CoRR*, abs/2004.04906, 2020. URL `https://arxiv.org/abs/2004.04906`.

Kelley J.F. *Cal-a natural language program developed with the oz paradigm: Implications for supercomputing systems*. IBM Thomas J. Watson Research Center, 1985.

Khattab O. and Zaharia M. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 39–48, 2020.

Kočiský T., Schwarz J., Blunsom P., Dyer C., Hermann K.M., Melis G., and Grefenstette E. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. URL `https://www.aclweb.org/anthology/Q18-1023`.

Kwiatkowski T., Palomaki J., Redfield O., Collins M., Parikh A., Alberti C., Epstein D., Polosukhin I., Devlin J., Lee K., *et al.*. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Lan Z., Chen M., Goodman S., Gimpel K., Sharma P., and Soricut R. Albert: A lite bert for self-supervised learning of language representations. *International Conference on Learning Representations*, 2020.

Lehnert W.G. A conceptual theory of question answering. *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1*, 158–164, 1977.

Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., and Zettlemoyer L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880, 2020.

Li J., Miller A.H., Chopra S., Ranzato M., and Weston J. Dialogue Learning With Human-in-the-Loop. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=HJgXCV9xx`.

Li Z., Sharma P., Lu X.H., Cheung J.C., and Reddy S. Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment. *arXiv preprint arXiv:2204.03025*, 2022.

Lin S.C., Yang J.H., Nogueira R., Tsai M.F., Wang C.J., and Lin J. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. *arXiv preprint arXiv:2004.01909*, 2020.

Lin S., Hilton J., and Evans O. TruthfulQA: Measuring How Models Mimic Human Falsehoods, 2021.

Liu B., Yu T., Lane I., and Mengshoel O.J. Customized Nonlinear Bandits for Online Response Selection in Neural Conversation Models. *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Liu S., Zhang X., Zhang S., Wang H., and Zhang W. Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9(18):3698, 2019a.

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., and Stoyanov V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b.

Lowe R., Pow N., Serban I.V., and Pineau J. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 285–294, 2015.

Madaan A., Tandon N., Gupta P., Hallinan S., Gao L., Wiegreffe S., Alon U., Dziri N., Prabhumoye S., Yang Y., *et al.*. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

Mishra A. and Jain S.K. A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28 (3):345–361, 2016.

Nguyen T., Rosenberg M., Song X., Gao J., Tiwary S., Majumder R., and Deng L. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268, 2016. URL `http://arxiv.org/abs/1611.09268`.

Ortega P.A., Kunesch M., Delétang G., Genewein T., Grau-Moya J., Veness J., Buchli J., Degrave J., Piot B., Perolat J., *et al.*. Shaking the foundations: delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819*, 2021.

Otegi A., Campos J.A., Azkune G., Soroa A., and Agirre E. Automatic evaluation vs. user preference in neural textual QuestionAnswering over COVID-19 scientific literature. *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.nlpcovid19-2.15`.

Otegi A., San Vicente I., Saralegi X., Peñas A., Lozano B., and Agirre E. Information retrieval and question answering: A case study on covid-19 scientific literature. *Knowledge-Based Systems*, 240:108072, 2022.

Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., *et al.*. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pan X., Sun K., Yu D., Chen J., Ji H., Cardie C., and Yu D. Improving question answering with external knowledge. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 27–37, 2019.

Peng B., Li C., Li J., Shayandeh S., Liden L., and Gao J. Soloist: Few-shot task-oriented dialog with a single pretrained auto-regressive model. *arXiv preprint arXiv:2005.05298*, 3, 2020.

Pennington J., Socher R., and Manning C.D. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543, 2014.

Perez E., Ringer S., Lukošiūtė K., Nguyen K., Chen E., Heiner S., Pettit C., Olsson C., Kundu S., Kadavath S., *et al.*. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., and Zettlemoyer L. Deep contextualized word representations. *Proceedings of NAACL-HLT*, 2227–2237, 2018.

Pradel C., Sileo D., Rodrigo Á., Peñas A., and Agirre E. Question answering when knowledge bases are incomplete. *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*, 43–54. Springer, 2020.

Puri R., Spring R., Shoeybi M., Patwary M., and Catanzaro B. Training question answering models from synthetic data. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5811–5826, 2020.

Qiu B., Chen X., Xu J., and Sun Y. A survey on neural machine reading comprehension. *arXiv preprint arXiv:1906.03824*, 2019.

Qu C., Yang L., Chen C., Qiu M., Croft W.B., and Iyyer M. Open-retrieval conversational question answering. *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 539–548, 2020.

Qu C., Yang L., Qiu M., Croft W.B., Zhang Y., and Iyyer M. BERT with history answer embedding for conversational question answering. *Proceedings of the*

42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 1133–1136, 2019.

Radford A., Narasimhan K., Salimans T., Sutskever I., *et al.*. Improving language understanding by generative pre-training. 2018. URL `https://openai-assets.s3.amazonaws.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I., *et al.*. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rae J.W., Borgeaud S., Cai T., Millican K., Hoffmann J., Song F., Aslanides J., Henderson S., Ring R., Young S., *et al.*. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. URL `https://arxiv.org/pdf/2112.11446.pdf`.

Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., and Liu P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Rajpurkar P., Jia R., and Liang P. Know what you don't know: Unanswerable questions for squad. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 784–789, 2018.

Rajpurkar P., Zhang J., Lopyrev K., and Liang P. Squad: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392, 2016.

Rasley J., Rajbhandari S., Ruwase O., and He Y. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3505–3506, 2020.

Reddy S., Chen D., and Manning C.D. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

Riloff E. and Thelen M. A rule-based question answering system for reading comprehension tests. *ANLP-NAACL 2000 workshop: reading comprehension tests as evaluation for computer-based language understanding systems*, 2000.

Robertson S., Zaragoza H., *et al.*. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Robertson S.E., Walker S., Jones S., Hancock-Beaulieu M.M., Gatford M., *et al.*. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

Roloff M.E. *Interpersonal communication*. Sage, 1981.

Salton G. The smart retrieval system: Experiments in automatic document processing. *IEEE Transactions on Professional Communication*, 17–17, 1972.

Sanh V., Webson A., Raffel C., Bach S., Sutawika L., Alyafeai Z., Chaffin A., Stiegler A., Raja A., Dey M., *et al.*. Multitask prompted training enables zero-shot task generalization. *International Conference on Learning Representations*, 2022.

Saunders W., Yeh C., Wu J., Bills S., Ouyang L., Ward J., and Leike J. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.

Scao T.L., Fan A., Akiki C., Pavlick E., Ilić S., Hesslow D., Castagné R., Luccioni A.S., Yvon F., Gallé M., *et al.*. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Scheurer J., Campos J.A., Chan J.S., Chen A., Cho K., and Perez E. Training language models with language feedback. *The First Workshop on Learning with Natural Language Supervision at ACL*, 2022.

Scheurer J., Campos J.A., Korbak T., Chan J.S., Chen A., Cho K., and Perez E. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*, 2023.

Schulman J., Wolski F., Dhariwal P., Radford A., and Klimov O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Seo M., Kembhavi A., Farhadi A., and Hajishirzi H. Bidirectional attention flow for machine comprehension. *International Conference on Learning Representations*, 2017.

Shakeri S., dos Santos C., Zhu H., Ng P., Nan F., Wang Z., Nallapati R., and Xiang B. End-to-end synthetic data generation for domain adaptation of question answering systems. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5445–5460, 2020.

Shi W., Dinan E., Shuster K., Weston J., and Xu J. When life gives you lemons, make cherryade: Converting feedback from bad responses into good labels. *arXiv preprint arXiv:2210.15893*, 2022.

Shoeybi M., Patwary M., Puri R., LeGresley P., Casper J., and Catanzaro B. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Snoek J., Larochelle H., and Adams R.P. Practical Bayesian Optimization of Machine Learning Algorithms. In Pereira F., Burges C.J.C., Bottou L., and Weinberger K.Q., editors, *Advances in Neural Information Processing Systems 25*, 2951–2959. Curran Associates, Inc., 2012. URL `http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf`.

Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

Stiennon N., Ouyang L., Wu J., Ziegler D., Lowe R., Voss C., Radford A., Amodei D., and Christiano P.F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Sutton R.S., Barto A.G., *et al.*. *Introduction to reinforcement learning*, 135 lib. MIT press Cambridge, 1998.

Thakur N., Reimers N., Rücklé A., Srivastava A., and Gurevych I. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *CoRR*, abs/2104.08663, 2021. URL `https://arxiv.org/abs/2104.08663`.

Thompson W.R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Thoppilan R., De Freitas D., Hall J., Shazeer N., Kulshreshtha A., Cheng H.T., Jin A., Bos T., Baker L., Du Y., *et al.*. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Trischler A., Wang T., Yuan X., Harris J., Sordoni A., Bachman P., and Suleman K. NewsQA: A machine comprehension dataset. *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-2623.

Urbizu G., San Vicente I., Saralegi X., Agerri R., and Soroa A. Basqueglue: A natural language understanding benchmark for basque. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1603–1612, 2022.

Vakulenko S., Longpre S., Tu Z., and Anantha R. Question rewriting for conversational question answering. *Proceedings of the 14th ACM international conference on web search and data mining*, 355–363, 2021.

Vinyals O., Fortunato M., and Jaitly N. Pointer networks. *Advances in neural information processing systems*, 28, 2015.

Wang A., Pruksachatkun Y., Nangia N., Singh A., Michael J., Hill F., Levy O., and Bowman S. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019a.

Wang A., Singh A., Michael J., Hill F., Levy O., and Bowman S.R. Glue: A multi-task benchmark and analysis platform for natural language understanding. *International Conference on Learning Representations*, 2019b.

Wang S. and Jiang J. Machine comprehension using match-lstm and answer pointer. *International Conference on Learning Representations*, 2017.

Weston J.E. Dialog-based language learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Davison J., Shleifer S., von Platen P., Ma C., Jernite Y., Plu J., Xu C., Le Scao T., Gugger S., Drame M., Lhoest Q., and Rush A. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45, Online, October 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-demos.6.

Xu C., Guo D., Duan N., and McAuley J. Laprador: Unsupervised pretrained dense retriever for zero-shot text retrieval. *Findings of the Association for Computational Linguistics: ACL 2022*, 3557–3569, 2022a.

Xu J., Ju D., Li M., Boureau Y.L., Weston J., and Dinan E. Bot-adversarial dialogue for safe conversational agents. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2950–2968, 2021.

Xu J., Ung M., Komeili M., Arora K., Boureau Y.L., and Weston J. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. *arXiv preprint arXiv:2208.03270*, 2022b.

Yang A., Wang Q., Liu J., Liu K., Lyu Y., Wu H., She Q., and Li S. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2346–2357, 2019.

Yang Z., Qi P., Zhang S., Bengio Y., Cohen W.W., Salakhutdinov R., and Manning C.D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

Yatskar M. A qualitative comparison of coqa, squad 2.0 and quac. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2318–2323, 2019.

Zhang S., Roller S., Goyal N., Artetxe M., Chen M., Chen S., Dewan C., Diab M., Li X., Lin X.V., *et al.*. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhao W.X., Liu J., Ren R., and Wen J.R. Dense text retrieval based on pretrained language models: A survey. *arXiv preprint arXiv:2211.14876*, 2022.

Glosategia

adostasun maila *agreement rate*

artearen egoera *state of the art*

aurrentrenatu *pretrain*

aurrerantza elikatzen den sare *feedforward network*

ausazko aldagai *random variable*

autonormalizazio *self-normalization*

azpilaginketa *subsampling*

batezbesteko doitasun *mean average precission*

bero mapa *heat map*

datu multzo *dataset*

doikuntza *fine-tuning*

doitasun *precission*

domeinu irekiko galdera-erantzun sistema *open domain question answering system*

domeinuz kanpo *out-of-domain*

elkarrizketa ekintza *dialogue act*

entropia gurutzatu *cross-entropy*

estaldura *recall*

gainbeiratu *supervised*

galdera-erantzunetan oinarritutako elkarrizketa *conversational question answering*

galdera-erantzun sistema *question answering system*

garrantzi bidezko laginketa *importance sampling*

garrantzi pisu *importance weight*

geruza anitzeko perzeptroi *multilayer perceptron*

geruza ezkutu *hidden layer*

goiz-eten *early stopping*

hitz-bektore *word embedding*

hizkuntza-eredu *language model*

helburu-distribuzio *target distribution*

informazio berreskurapena *information retrieval*

komunitate galdera-erantzun gune *community question answering site*

laginketa *sampling*

lagin espazioa *sample space*

makina bidezko irakurketa ulermen *machine reading comprehension*

maiz egiten diren galderak *frequently asked questions*

modeloaren sinesmen *model's belief*

optimizazio bayestiar *bayesian optimization*

proposamen distribuzio *proposal distribution*

sailkapen geruza *classification layer*

sistema gainbegiratu *supervised system*

A.1 Original papers

In this appendix we present the original papers presented in the manuscript of this thesis in the recommended reading order.

DoQA - Accessing Domain-Specific FAQs via Conversational QA

Jon Ander Campos[1], **Arantxa Otegi**[1], **Aitor Soroa**[1],
Jan Deriu[2], **Mark Cieliebak**[2], **Eneko Agirre**[1]
[1]University of the Basque Country (UPV/EHU)
[2]Zurich University of Applied Sciences (ZHAW)
[1]{jonander.campos, arantza.otegi, e.agirre, a.soroa}@ehu.eus
[2]{jan.deriu, mark.cieliebak}@zhaw.ch

Abstract

The goal of this work is to build conversational Question Answering (QA) interfaces for the large body of domain-specific information available in FAQ sites. We present DoQA, a dataset with 2,437 dialogues and 10,917 QA pairs. The dialogues are collected from three Stack Exchange sites using the Wizard of Oz method with crowdsourcing. Compared to previous work, DoQA comprises well-defined information needs, leading to more coherent and natural conversations with less factoid questions and is multi-domain. In addition, we introduce a more realistic information retrieval (IR) scenario where the system needs to find the answer in any of the FAQ documents. The results of an existing, strong, system show that, thanks to transfer learning from a Wikipedia QA dataset and fine tuning on a single FAQ domain, it is possible to build high quality conversational QA systems for FAQs without in-domain training data. The good results carry over into the more challenging IR scenario. In both cases, there is still ample room for improvement, as indicated by the higher human upperbound.

1 Introduction

The overarching objective of our work is to access the large body of domain-specific information available in Frequently Asked Question sites (FAQ for short) via conversational Question Answering (QA) systems. In particular, we want to know whether current techniques are able to work with limited training data, and without needing to gather data for each target FAQ domain. In this paper we present **DoQA**, a task and associated dataset for accessing domain-specific FAQs via conversational QA[1]. The dataset contains 2,437 information-seeking question/answer dialogues on three different domains

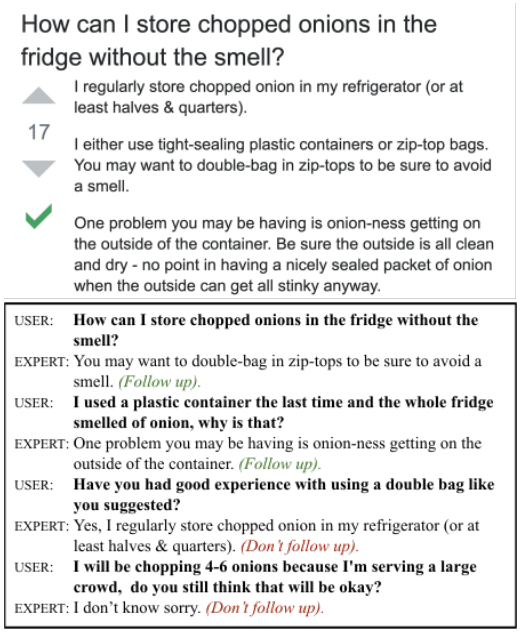


Figure 1: A dialogue about cooking. On top, the original post, comprising a topic and an excerpt of the answer passage. In italics, dialogue acts (cf. Section 3).

(10,917 questions in total). These dialogues are created using the Wizard of Oz technique by crowdworkers that play the following two roles: the **user** asks questions about a given topic posted in Stack Exchange[2], and the **domain expert** replies to the questions by selecting a short span of text from the long textual reply in the original post. The first question is prompted by the real FAQ question, which sets the topic of interest driving the user questions. In addition to the extractive span, we also allow experts to rephrase it, in order to provide an abstractive, more natural, answer. The dataset covers unanswerable questions and some relevant dialogue acts. We focused on three different domains: Cooking, Travel and Movies. These forums

[1]The DoQA dataset is available here: http://ixa.eus/node/12931

[2]https://stackexchange.com/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7302–7314
July 5 - 10, 2020.

are some of the most active ones and contain knowledge of general interest, making it easily accessible for crowdworkers. DoQA contains two scenarios: in the standard scenario the test data comprises the questions and the target document from which the answers need to be extracted; in the information retrieval (IR) scenario the test data contains the questions, but the target document is unknown, and the system needs to select the documents which contain the answers among all documents in the collection.

Previous work on conversational QA datasets include CoQA (Reddy et al., 2018) and QuAC (Choi et al., 2018). The main focus of CoQA are reading comprehension questions, which are produced with access to the target paragraph. The topic of the questions are delimited by the paragraph, which leads to specific questions about details in the paragraph. Choi et al. (2018) observed that a large percentage of CoQA answers are named entities or short noun phrases. In QuAC, the topic of the conversation is set by a title and first paragraph of a Wikipedia article about people. The user makes up questions about the person of interest. Note that, contrary to our setting, there is no real information need in any of those datasets, which can lead to less coherent conversations: any question about the paragraph or person of interest is valid, respectively.

DoQA makes the following **contributions**. Firstly, contrary to made-up reading comprehension tasks, DoQA reflects real user needs, as defined by a topic in an existing FAQ. Good results on DoQA are of practical interest, as they would show that effective conversational QA interfaces to FAQs can be built. Secondly, for the same reason, the conversations in DoQA are more coherent, natural and contain less factoids than other datasets, as shown by our analysis. Thirdly, the IR scenario and the multiple domains make DoQA more challenging and realistic. Table 1 summarizes the characteristics of DoQA.

Although one could question the small size of our dataset, our goal is to test whether current techniques are able to work with limited training data, and without needing to gather data for each target FAQ domain. We thus present results of an existing strong conversational QA model with limited and out-of-domain data. The system trained on Wikipedia data (QuAC) provides some weak results which are improved when fine-tuning on the FAQ dataset. Our empirical contribution is to show that a relatively low amount of training in one FAQ dataset (1000 dialogues on Cooking) is sufficient for strong results on Cooking (comparable to those obtained in the QuAC dataset with larger amounts of training data), but also on two other totally different domains with no in-domain training data (Movies and Travel). In all cases scores over 50 F1 are reported. Regarding the IR scenario, an IR module complements the conversational system, with a relatively modest drop in performance. The gap with respect to human performance is over 30 points, showing that there is still ample room for system improvement.

	DoQA	QuAC	CoQA
Real information need	▲		
Naturalness	▲		
Dialogue coherence	▲		
Non-factoid questions	▲		
Unanswerable questions	▲	▲	
Dialogue acts	▲	▲	
Multi-domain	▲		▲
IR scenario	▲		

Table 1: Summary of the characteristics of DoQA compared to QuAC and CoQA. ▲ for positive.

2 Related Work

Conversational QA systems stem from the body of work on Reading Comprehension, whose goal is to test the capacity of a system to understand a document by answering any question posed over its content. Recent work on the field has resulted in the creation of multiple datasets (Rajpurkar et al., 2016; Trischler et al., 2017; Nguyen et al., 2016; Kočiský et al., 2018; Dunn et al., 2017). These datasets are typically composed of multiple question/answer pairs, often along with a reference passage from which the answer is curated. Whereas the questions are always in free text form, some datasets represent the answers as a contiguous span in the reference passage, while others contain free form answers. The former are usually referred as *extractive*, whereas the latter are called *abstractive*. All in all, in these QA datasets the queries are unrelated to each other, and thus there is no dialogue structure involved.

Iyyer et al. (2017) propose to answer complex queries by decomposing them into sequences of single, co-referent queries. The question sequence can be seen as different turns in a dialogue, and each question refers and refines previous ones. The

authors present the SequentialQA dataset, which comprises 6K question sequences posed over the content of Wikipedia tables. In the case of our task, it is the user who makes several questions in sequence.

More similar to our work, CoQA (Reddy et al., 2018) and QuAC (Choi et al., 2018) are two conversational QA datasets comprising QA dialogues that fulfill the information need of a user by answering questions about different topics. Similarly to our, both datasets are built by crowdsourcing, where one person (the questioner) is presented with a topic and has to pose free-form questions about it. Another person (the answerer) has to select an answer to the question by choosing an excerpt from the relevant passage describing the topic. Some of the questions in both datasets are unanswerable, and access to previous questions and answers are needed in order to answer some of the questions.

CoQA contains 127k questions with answers, obtained from 8k conversations about passages from broad domains, ranging from children stories to science. The answers are also excerpts from the relevant passage, but answerers have the choice of reformulating them. The authors report that 78% of the answers had at least one edit. Although reformulating answers can yield to more natural dialogues, Yatskar (2018) showed that span based systems can in principle obtain a performance up to 97.8 points F1, showing that editing the answers does not yield to systems with better quality. In CoQA, both questioner and answerer have access to the full passage, which guides the conversation towards the specific information conveyed in it.

QuAC is a dataset that contains 14k information-seeking question answering dialogues. The dialogues in QuAC are about a specific section in Wikipedia articles about people. The answerer has access to the full section text, whereas the questioner only sees the section's title and the first paragraph of the main article, which serves as inspiration when formulating the queries. QuAC also contains dialogue acts in each turn, which are useful when collecting the dialogues, as they can be used by the answerer to indicate to questioner whether to continue making questions about the last answer or drift to other aspects of the topic. We will compare CoQA and QuAC in more detail in Section 4.

Previous conversational QA datasets provide the relevant document or passage that contain the answer of a query. However, in many real world scenarios such as FAQs, the answers need to be searched over the whole document collection. In related question answering research, Chen et al. (2017) and Watanabe et al. (2017) combine retrieval and answer extraction on a large set of documents. In (Talmor and Berant, 2018) the authors propose decomposing complex questions into a sequence of simple questions, and using search engines to answer those single questions, from which the final answer is computed. We find that requiring the system to search for relevant documents and passages is more realistic, and DoQA is the first conversational QA task incorporating this scenario.

In contemporary work, Castelli et al. (2019) present a question answering dataset for the technical support domain which focuses on actual questions posed by users and has a real-world size with only 600 training instances. It also requires systems to examine 50 documents per query. Our work has similar motivations for setting up more realistic tasks, and is complementary in the sense that we cover non-technical domains and conversatioal QA.

Community Question Answering has been also the focus of two related tasks (Nakov et al., 2016, 2017), where, given a new question and a collection of pre-existing questions and answers, the systems need to rank the answers that are most useful for answering the new question.

3 Dataset Collection

This section describes our conversational QA dataset collection process which consists of an interactive task designed for two crowdworkers in Amazon Mechanical Turk (AMT).

3.1 FAQ Post Selection

We collected topic-answer pairs for the three different domains from the Stack Exchange data dumps. We focused on the Cooking[3], Travel[4] and Movies[5] domains, as they are active forums and contain knowledge of general interest, making it easily accessible and attractive for crowdworkers. Note that the posts in Stack Exchange (as in most FAQ sites) comprise broad questions which often require lengthy answers. We refer to the question in the post as *topic* and to the long answer in the post as *passage* (not to be confused with the actual ques-

[3] https://cooking.stackexchange.com/
[4] https://travel.stackexchange.com/
[5] https://movies.stackexchange.com/

tions/answers in the collected dialogues). Figure 1 shows an example of a topic and its corresponding passage for the Cooking domain. More details on post filtering and selection can be found in Appendix A.

3.2 Crowdsourcing Task

For the annotation process, we defined a HIT in AMT as the task of generating a dialogue about a specific topic between two workers (the specifications of the defined HIT can be found in Appendix B). One of the workers (the user) asks questions to the second one (the domain expert) about a certain topic from a Stack Exchange Cooking, Travel or Movies thread. The worker who adopts the user role has access to a small paragraph that introduces the topic. Having this information, he must ask free text questions. The first question of every dialogue must be the title of the topic that appears in the title of the Stack Exchange thread. The domain expert has access to the whole answer passage and he/she answers the query by selecting a span of text from it. In order to make the dialogue look more natural, the domain expert has the opportunity to edit the answer, but note that if he does so the answer will not match the content of the text span anymore. Therefore, and following Yatskar (2018), we motivate minimal modifications by copying the selected text span directly into the answer field in the web application. In addition to the span of text, the expert has to give feedback to the user with one of the following dialogue acts: an affirmation act, which is is required when the question is a Yes/No question (*yes*, *no* or *neither*); an answerability act, which defines if the question has an answer or not (*answerable* or *no answer*). When no answer is selected, the returned string is "I don't know"; and a continuation dialogue act, which is used for leading the user to the most interesting topics (*follow up* or *don't follow up*). The last dialogue act is used to minimally guide the user in his/her questions, where the expert can encourage (or dicourage) the user to continue with questions related to his last questions using *follow up* (or alternatively *don't follow up*). These dialogue acts are the same as in QuAC, but we discarded the *maybe follow up* act from the continuation act because we felt it was not intuitive enough.

Dialogues are ended when a maximum of 8 question and answer pairs is reached, when 3 unanswerable questions have been asked, or when 10 minutes time limit is reached. The purpose of these limits is to avoid long and repetitive dialogues, because real threads of the selected domains are very focused on a certain topic. Dialogues are only accepted if they have a minimum length of 2 question and answer pairs and if they have at least one answer that is not "I don't know sorry".

The data collection interface is based on CoCoA[6], which we modified. The interfaces for the user and expert are shown in Appendix C.

	Cooking			Travel	Movies
	Train	Dev.	Test	Test	Test
Questions	4,612	911	1,797	1,713	1,884
Dialogues	1,037	200	400	400	400
Unique sections	546	162	400	400	400
Tokens / question	10.79	10.14	10.66	10.45	9.45
Tokens / answer	13.19	13.10	12.58	13.47	12.40
Dialogue turns	4.47	4.55	4.49	4.28	4.71
Extractive %	69.68	67.18	66.95	65.44	74.15
Abstractive %	30.32	32.82	33.05	34.56	25.85
Yes/No %	20.22	21.07	22.20	25.10	18.05
I don't know %	27.55	27.33	29.71	22.83	29.41

Table 2: Statistics of the different domains of DoQA.

3.3 Dataset Details

Following usual practice, we divided the main Cooking dataset into a train, development and test splits. For the other two domains, Travel and Movies, we only have the test split. Statistics for all the domains and splits are shown in Table 2.

The splits of the Cooking dataset have very similar characteristics, so we can expect them to be valid representatives of the whole Cooking dataset. In the test splits we do not allow more than one dialogue about the same section, as it can end up producing inaccurate evaluation of the models.

3.4 Collecting Multiple Answers

In order to estimate the performance of a human in the task, we collected additional answers for the test splits for the three domains in a second round, after having completed the dialogues. For each question in the dialogues collected in the first round, we show to the worker the previous questions and answers in the dialogue (if available), and he has to provide an answer span. The interface for the collection of multiple answers can be seen in Appendix D.

[6] https://github.com/stanfordnlp/cocoa (He et al., 2017)

3.5 Information Retrieval Scenario

In the usual setting for this kind of tasks, the system is given the question and the passage where the answer is to be extracted from. In a realistic scenario, however, relevant answer passages that may contain the answer will need to be retrieved first. More specifically, if a user has an information need and asks a question to a conversational QA system on a FAQ, the system can search for similar questions which have already been answered, or the system can directly search in existing answer passages. In other words, there are two ways to check automatically if the forum contains a relevant answer passage to a new question: (1) question retrieval, where relevant or similar questions are searched (and thus, the answer for this relevant question is taken as a relevant answer), and (2) answer retrieval, where relevant answers are searched directly among existing answers.

We added information about both relevant cases to the main Cooking dataset, in the form of the 20 most relevant answer passages for each dialogue in the dataset. We followed a basic approach to get these relevant answer passages. We created two separate indexes using an IR system[7] for the two mentioned approaches, question and answer retrieval. For the former, we indexed the original topics posted in the forum; and for the latter, we indexed the answer passages for each post in the forum. Then, for each dialogue in the development and test splits, the top 20 documents were retrieved using the first question of the dialogue. Given that the dialogues are about a single topic, we only use the first question in the dialogue, and then use the retrieved passages for the rest of questions in the dialogue as well.

The question retrieval approach yields very good results (0.94 precision at one), as expected, as the crowdworker doing the questions has access to the topic when asking the first question and usually did minor edits. The results for answer retrieval are more modest, 0.54 precision at one. The results section shows the results of the conversational QA system when relying on the passages returned by the IR module.

4 Dataset Analysis

Overall statistics In this section we present an quantitative and qualitative analysis of DoQA and we compare them to similar conversational datasets like QuAC and CoQA, stressing its similarities and differences.

	DoQA	QuAC	CoQA
Questions	10,917	98,407	127,000
Dialogues	2,437	13,594	8,399
Tokens / question	10.43	6.5	5.5
Tokens / answer	12.99	14.6	2.7
Dialogue turns	4.48	7.2	15.2
Extractive %	69.13	100	66.8
Abstractive %	30.87	-	33.2
Yes/No %	21.01	25.8	-
I don't know %	27.47	20.3	1.3

Table 3: Statistics of DoQA compared to QuAC and CoQA.

Table 3 shows the overall statistics of DoQA, together with the statistics of QuAC and CoQA. As can be seen, DoQA has the smallest amount of questions and dialogues. However, other features makes it very interesting for the research of conversational QA. For instance, the average tokens per questions and answers (10.43 and 12.99, respectively) are closer to real dialogues if we compare to the other datasets. Specially CoQA has very short questions and answers on average, suggesting that CoQA is closer to factoid QA than dialogue, as human dialogues tend to be longer and convoluted, not just short answers. DoQA has the lower ratio of questions per dialogue, which is expected, as most of the dialogues are about a very specific topic and the user is satisfied and gets the answer without the need of long dialogues. CoQA ends up on having almost all of its questions answerable, facing the same issues as SQuAD 1.0 (Rajpurkar et al., 2016) that motivated the addition of unanswerable questions in SQuAD 2.0 (Rajpurkar et al., 2018).

We also have the results of a short survey that workers had to respond to at the end of each HIT. On the one hand, the user had to give feedback on how satisfied was with the answers of the expert in a scale of 1-5. The average satisfaction was 3.9. On the other hand, the expert had to give feedback on how sensible were the questions and the helpfulness of the answers. The average scores obtained were 4.27 and 4.10, respectively, which makes the AMT task satisfactory.

Naturalness One of the main positive aspects of our dataset is the naturalness of the dialogues that other similar datasets like QuAC do not have. The answers of DoQA come from a forum where the answer text is directed to a person who posted the

[7] Solr https://lucene.apache.org/solr/

question, and does not come from a much formal text like Wikipedia, as it is the case of QuAC. The naturalness and casual register of the former it is more adequate than the formal register of the latter for a conversational QA system. The dialogue in Figure 1 is a clear example of such naturalness, where the expert answers to the user with casual and directed expressions like "*You may want*" and "*you may be having*". To verify whether dialogues in DoQA are more natural than the ones in QuAC, we sample randomly 50 dialogues in DoQA Cooking domain and QuAC and performed A/B testing to determine which of the two dialogues is more natural. This test showed that 84% of the times DoQA dialogues are more natural.

This naturalness is probably caused because a dialogue in DoQA is started by a user with a very specific aim or topic to solve in mind, and thus, follow-up questions are very related to previous answers, and all the questions are set within a context. In contrast, dialogues in QuAC do not show so clear objective and questions seem to be asked randomly. Dialogues in DoQA are ended when the initial information need of the user is satisfied and this adds naturalness to dialogues.

Further analysis of the samples showed that answers in DoQA seem to be more spontaneous because they have more orality aspects, such as higher level of expressivity ("*Normally when I try they end up burned not crispy!*", "*My biggest worry here would be...*", "*hey let's not be hasty*"), opinions ("*I came across a suggestion to cover the lid...*", "*I'd recommend simply adding...*", "*It sounds like fermentation to me*") and humor ("*well yeah but booze is booze*"). Contrarily, answers in QuAC are more hermetic and do not show any features of orality or spontaneity that a dialogue should have. All these features make DoQA dialogues look more natural.

We also analyzed the remaining 16% cases where DoQA dialogues appear less natural. In most of these dialogues there were responses that did not really answer the question. The following question (Q) and answer (A) pairs are good examples of it: (Q) "*Is the taste going to be significantly different?*" (A) "*there is cornstarch in confectioner's sugar*"; (Q) "*how about reheating?*" (A) "*When you defrost it, do so in your fridge leaving it overnight so that it defrosts gradually*"; (Q) "*Can I use my potatoes or carrots if they already have some roots?*" (A) "*The green portions of a potato are toxic*". In some of these cases the correct answer for the respective question is not in the answer text provided to the expert. If this was the case, the expert should answer "I don't know", instead of giving a nonsense answer.

Question types Table 4 includes the most frequent two initial words of the questions in the Cooking dataset along with their percentages of occurrences and some examples. Most of the questions start with *what* and *how* (16.6% and 15.1% of the questions, respectively), which are also the most frequent in QuAC and CoQA. Contrary to them, the questions in the Cooking dataset do not refer to factoids, with the exception of "How long" questions. The questions in DoQA require long and complex answers. In contrast to this, in CoQA and QuAC many of the most frequent initial words such as *who*, *where*, and *when* indicate factoid questions. In order to confirm this fact, we manually inspected 50 random questions from the Cooking domain and QuAC datasets. This analysis revealed that 66% of the questions are non-factoid in the DoQA Cooking domain, showing that most of the questions are open-ended. These amount is larger than in QuAC, as in our analysis for QuAC we found that only 36% of the questions are non-factoid. These values differ slightly from those reported by Choi et al. (2018), as they say that about half of questions are non-factoid.

Context or history dependence The manual analysis also shows that 61% of the questions are dependent on the conversation history, as many questions have coreferences to previous questions or answers in the dialogue. For example, "*What are other methods to sharpen a knife?"*, "*How long should I cook it in the microwave?"*, "*Can you explain the science behind this cooking procedure?"*. Moreover, we could note that less than 1% ask further advice or tips about the current topic, confirming that these conversations are about specific topics where the user is satisfied with the expert answers after a few questions.

Dialogue coherence Related to the just mentioned fact that the user does not usually ask any other tips, users in DoQA do not tend to switch topics in a dialogue. In order to confirm it, we performed another A/B testing to the same 50 dialogues samples of the DoQA Cooking domain and QuAC to determine which of the two dialogues is more coherent, that is, which dialogue has a smoother flow. This test revealed that in 64% of

Bigram prefix		%	Example
What	is	30.8	What is the purpose of adding water to an egg wash?
(16.6%)	are	8.0	What are other methods to sharpen a knife?
How	do	24.0	How do you properly defrost frozen fish?
(15.1%)	long	21.9	How long should I cook it in the microwave?
Is	there	52.8	Is there a special tool available for cracking open a pistachio?
(10.5%)	it	19.8	Is it safe to cook with rainwater?
Do	you	70.7	Do you have any advice for storing green onions?
(7.6%)	I	16.1	Do I have to peel the apples?
Can	I	52.8	Can I put them back in the oven to reheat?
(5.5%)	you	25.3	Can you explain the science behind this cooking procedure?
I	have	19.6	I have been told that frying it would make it tastier, but is it healthier to grill or fry?
(5.0%)	am	15.3	I am cooking for somebody who doesn't eat shellfish, so is the fish sauce safe?
Why	is	22.1	Why is it important to increase the fermentation time?
(3.5%)	does	21.7	Why does my custard pudding taste like raw eggs?

Table 4: Most frequent initial words and bigrams in questions (Cooking domain).

the cases dialogues of DoQA are more coherent than QuAC. Only in 10% of the cases dialogues of DoQA are less coherent, with the remaining 26% equally coherent. We analyzed the 10% and saw that they contain similar questions one after the other, or repeated answers in the same dialogue.

Summary Table 1 summarizes the positive characteristics of DoQA compared to the similar datasets like QuAC and CoQA.

5 Task Definition

Given a textual passage and a question, traditional QA systems find an answer to the question within the passage. Conversational QA systems are more complex, as they need to deal with a sequence of possibly inter-dependent questions. That is, the meaning of the current question may depend on the dialogue history. For this reason, a dialogue history comprised by previous question/answer pairs is also provided to the system. In addition, some dialogue acts have to be predicted as an output: yes/no answers, which are required for affirmation questions, and continuation feedback, which might be useful for information-seeking dialogues.

We denote the answer passage as p, the dialogue history of questions and respective ground truth answers as $\{q_1, a_1, ...q_{k-1}, a_{k-1}\}$, current question as q_k, the answer span a_k which is delimited by its starting index i and ending index j in the passage p, and dialogue act list v. The dialogue act list contains $\{$*yes,no,-*$\}$ values for predicting affirmation and $\{$*follow-up,don't follow-up*$\}$ for continuation feedback.

6 Baseline Models

We present two strong baseline models to address our task. Although the state-of-the-art evolves quickly, our choice has the benefit of simplicity and strong performance.

BERT We took the fine-tuning approach for QA of BERT, which predicts the indexes i and j of the a_k answer span given p and q_k as input. This baseline has shown strong performance on QA datasets such as SQuAD (Devlin et al., 2018).

BERT+HAE The previous baseline does not model dialogue history. We used BERT with History Answer Embedding (HAE) as proposed by Qu et al. (2019) as a baseline that deals with the multi-turn problem, as this is the publicly available system that performs best in the QuAC leaderboard[8]. The system introduces dialogue history $\{q_1, a_1, ...q_{k-1}, a_{k-1}\}$ to BERT by adding a history answer embedding layer, which learns whether a token is part of history or not.

7 Evaluation

Evaluation metrics Given the similarity between QuAC and DoQA, we use the same evaluation metrics and criteria used in QuAC. F1 is the main evaluation metric and is computed by the overlap at word level of the prediction and reference answers. As the test set contains multiple answers for each question we take the maximum F1 among them. Note that when computing F1 QuAC filters

[8] accessed on August 20, 2019

		Cooking			Travel			Movies		
Setting	Model	F1	HEQ-Q	F1all	F1	HEQ-Q	F1all	F1	HEQ-Q	F1all
Native	BERT	40.1	35.1	38.3	36.2	34.8	34.8	36.1	33.5	35.0
	BERT+HAE	47.8	43.0	45.9	44.0	37.4	42.9	42.8	37.1	41.9
Zero-shot	BERT	40.2	34.7	38.9	34.0	30.1	33.1	38.2	33.2	37.4
	BERT+HAE	46.2	42.0	44.5	42.7	37.1	42.3	45.4	41.4	44.8
Transfer	BERT	43.3	37.8	42.4	40.6	33.6	40.1	41.8	36.3	41.3
	BERT+HAE	53.2	**48.3**	51.4	50.8	42.1	50.6	51.6	44.3	51.5
Transfer	BERT	43.1	37.0	42.0	40.6	33.4	40.5	42.0	34.5	41.6
all	BERT+HAE	**53.4**	46.9	**52.7**	**51.6**	**43.3**	**50.9**	**52.1**	**45.2**	**51.7**
Human		-	100.0	86.6	-	100.0	87.4	-	100.0	88.8

Table 5: Results of the baseline systems in the three DoQA domains (columns) in all four settings (rows). See text for explanation of each row. Note that Travel and Movies results are obtained without any Travel or Movies training data.

out answers with low agreement among human annotators. An additional F1-all is provided for the whole set. We also report HEQ-Q (human equivalence score on a question level) which measures the percentage of questions for which system F1 exceeds or matches human F1.

Experimental Setup We carried out experiments using the extractive information in DoQA, leaving the abstractive information for the future. The parameters we used to train the baseline models are the ones proposed in the original papers. We tested the models in four settings. In the **native** setting the Cooking DoQA train and dev data are used, the first for training and the second for early stopping. In the **zero-shot** setting we use QuAC training data for training and early stopping. In the **transfer** setting we use QuAC and Cooking DoQA for training. Finally, in the **transfer all** setting we additionally use the test data from the other two domains for training.

We also experimented on the IR scenario, using the provided IR rankings (see Section 3.5), which contain the top 20 passages for each dialogue. In the first experiment, *Top-1*, we just use the top 1 passage and apply the baseline BERT model. In a second experiment, *Top-20:BERT*, the passages are fed to the BERT model and the passage that contains the answer with highest confidence score is selected. Note that we discard passages that produced "I don't know" type of answers. In a third experiment, *Top-20:BERT*IR*, we select the passage with highest combined score according to BERT and the search engine.

All the reported results have been achieved using the BERT Base Uncased model.

Results Table 5 summarizes our results. In the bottom row we give the human upperbound. The three metrics used for evaluation behave similarly, so we focus on one (e.g. F1) for easier discussion. We report all three for completion and easier comparison with related datasets. In all settings and domains the BERT+HAE model yields better results than BERT, showing that **DoQA is indeed a conversational dataset**, where question and answer history needs to be modelled.

Regarding the different settings, we first focus on the **Cooking** dataset. The native scenario and the zero-shot settings yield similar results, showing that the 1000 dialogues on Cooking provide the same performance as 13000 dialogues on Wikipedia from QuAC[9]. The combination of both improves performance by 7 points ("Transfer" row), with small additional gains when adding Movies and Travel dialogues for further fine-tuning ("Transfer all" row). Note that the performance obtained for Cooking in the "Transfer" or "Transfer all" setting is **comparable to the one reported for QuAC**, where the training and test are from the same domain[10].

Yet, the most interesting results are those for the **Travel and Movies** domains, which do not have access to in-domain training data on Travel or Movies. In this case, the native and **transfer results with**

[9] When randomly subsampling QuAC to the same size as DoQA the results on the cooking domain fall down to 36.5.

[10] BERT+HAE obtains 62.4 in QuAC (Qu et al., 2019), 9 points higher than in DoQA Cooking, but note that QuAC contains more reference answers per question than DoQA, and thus the resulting F1 scores are higher. When evaluating BERT+HAE using a single reference answer in both datasets, the score is 45.9 on QuAC and 47.8 on the Cooking dataset of DoQA.

Model	F1	HEQ-Q	F1-all
Answer retrieval			
Top-1	**37.2**	**33.3**	**35.8**
Top-20:BERT	32.7	29.6	31.0
Top-20:BERT*IR	36.1	32.9	34.4
Question retrieval			
Top-1	**42.2**	**36.76**	**41.1**
Top-20:BERT	35.8	31.2	34.3
Top-20:BERT*IR	41.6	36.4	40.5

Table 6: Results on the IR scenario (Cooking domain). See text for explanation

no in-domain training are as high as those for Cooking. These results show that it is not necessary to train for each domain in a FAQ, and that training data from other FAQ domains is highly reusable.

The results obtained on out-of-domain test conversations (Movie and Travel) when trained on Wikipedia and Cooking are striking, as they are comparable to the in-domain results obtained for the Cooking test conversations. We hypothesize that when people write the answer documents in FAQ websites such as Stackexchange, they tend to use linguistic patterns that are common across domains such as Travel, Cooking or Movies. This is in contrast to Wikipedia text, which is produced with a different purpose, and might contain different linguistic patterns. As an example, in contrast to FAQ text, Wikipedia text does not contain first-person and second-person pronouns. We leave an analysis of this hypothesis for the future.

Table 6 presents the results of the experiments on the **IR scenario**. The simplest Top-1 approach is the best performing for both question and answer retrieval strategies. We leave the exploration of more sophisticated techniques for future work. The results using question retrieval are very close to those in Table 5. Given the large gap in the IR results in Section 3.5 for answer retrieval, it is a surprise to see a small 5 point decrease with respect to question retrieval. We found that there is a high correlation between the errors of the dialogue system and the answer retrieval system, which explains the smaller difference. In both retrieval strategies the **results are close** to the performance obtaining when having access to the reference target passage.

8 Conclusion and Future Work

The goal of this work is to access the large body of domain-specific information in the form of Frequently Asked Question sites via conversational QA systems. We have presented DoQA, a dataset for accessing Domain specific FAQs via conversational QA that contains 2,437 information-seeking dialogues on the Cooking, Travel and Movies domain (10,917 questions in total). These dialogues are created by crowdworkers that play the following two roles: the user asks questions about a certain topic posted in Stack Exchange, and the domain expert who replies to the questions by selecting a short span of text from the long textual reply in the original post. The expert can rephrase the selected span, in order to make it look more natural. In contrast to previous conversational QA datasets, our dataset responds to a real information need, is multi-domain, more natural and coherent. DoQA introduces a more realistic scenario where the passage with the answer needs to be retrieved.

Together with the dataset, we presented results of a strong conversational model, including transfer learning from Wikipedia QA datasets to our FAQ dataset. Our dataset and experiments show that it is possible to access domain-specific FAQs using conversational QA systems with little or no in-domain training data, yielding quality which is comparable to those reported in QuAC.

For the future, we would like to exploit the abstractive answers in our dataset, explore more sophisticated systems in both scenarios and perform user studies to study how real users interact with a conversational QA system when accessing FAQs.

Acknowledgments

This research was partially supported by a Google Faculty Award, EU ERA-Net CHIST-ERA LIHLITH funded by the Agencia Estatal de Investigación (AEI, Spain) project PCIN-2017-118 and the Swiss National Science Foundation (SNF, Switzerland) project 20CH21 174237, project DeepReading (RTI2018-096846-BC21) supported by the Ministry of Science, Innovation and Universities of the Spanish Government, the Basque Government (DL4NLP KK-2019/00045 and excellence research group), project BigKnowledge (Ayudas Fundación BBVA a Equipos de Investigación Científica 2018) and the NVIDIA GPU grant program. Jon Ander Campos enjoys a doctoral grant from the Spanish MECD.

References

Vittorio Castelli, Rishav Chakravarti, Saswati Dana, Anthony Ferritto, Radu Florian, Martin Franz, Dinesh Garg, Dinesh Khandelwal, Scott McCarley, Mike McCawley, Mohamed Nasr, Lin Pan, Cezar Pendus, John Pitrelli, Saurabh Pujar, Salim Roukos, Andrzej Sakrajda, Avirup Sil, Rosario Uceda-Sosa, Todd Ward, and Rong Zhang. 2019. The TechQA Dataset.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. *CoRR*, abs/1704.05179.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776.

Mohit Iyyer, Wen tau Yih, and Ming-Wei Chang. 2017. Search-based Neural Structured Learning for Sequential Question Answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. Semeval-2017 task 3: Community question answering. *arXiv preprint arXiv:1912.00730*.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 525–545, San Diego, California. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv, abs/1611.09268*.

Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. BERT with History Answer Embedding for Conversational Question Answering. *CoRR*, abs/1905.05412.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. *arXiv preprint arXiv:1806.03822*.

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

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

Alon Talmor and Jonathan Berant. 2018. The Web as a Knowledge-Base for Answering Complex Questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Yusuke Watanabe, Bhuwan Dhingra, and Ruslan Salakhutdinov. 2017. Question Answering from Unstructured Text by Retrieval and Comprehension. *CoRR*, abs/1703.08885.

Mark Yatskar. 2018. A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. *arXiv preprint arXiv:1809.10735*.

A FAQ Post Selection

First, we downloaded the data dumps from September 2018 for cooking forum and September 2019 for travel and movies forums. We then removed threads with unaccepted answers. At this point we did a preliminary analysis of the cooking topic scores and the lengths of the answer passages. Regarding the scores, we realized that all topic scores were in the range $[-6, 240]$. After manually analysing some random samples, we concluded that even low scoring topics had a good quality, except for the ones with negative scores. Regarding the length of the answer passages, some of them were too long for our task (up to $2,960$ tokens), as very long passages makes the task very tedious. Taking all this into account, we applied the following filters to the topic-passage pairs for the three domains:

- Topics with score $<= 0$ are removed, as we are not interested in badly asked questions.
- Topic titles with more than one question mark are removed. The reason behind this filter is that we are interested in having the topic titles as the first question of our dialogues and we are not interested in having more than one question per dialogue turn.
- The length of the answer passage has to be greater than 50 and shorter than 250 tokens. This way, we try to ensure that the answer passage is long enough for collecting dialogue, but not too long for avoiding tedious answer spotting.
- Answers that contain HTML tags such as hyperlinks, images, code, etc. are removed.

B Amazon Mechanical Turk HIT Specifications

In order to select the workers in AMT, we defined the HIT with the following specifications:

- HIT approval rate $\geq 98\%$.
- Approved HITs ≥ 1000
- Location of the workers: English speaking countries.

We paid the workers $\$0.10$ for doing the HIT and a bonus of $\$0.33$ for each question or answer given during the task except for the "I don't know sorry" case where $\$0.05$ was paid. This difference in the payment motivates the workers to force themselves to find the actual answer in the passage, because answering "I don't know" is less demanding than searching for the correct answer span. The average price for each dialogue is $\$3.2$.

C Dialogue Collection Interfaces

For dialogue collection, the worker carrying out the user role used the interface shown in Figure 2 and one with the expert role used the interface displayed in Figure 3.

D Multiple Answers Collection Interface

The interface used for multiple answers collection can be seen in Figure 4.

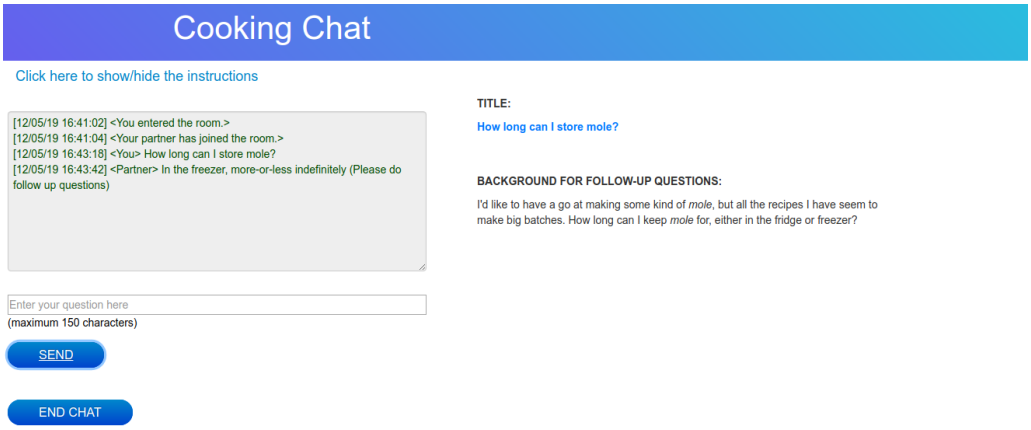


Figure 2: Dialogue collection interface for the user.

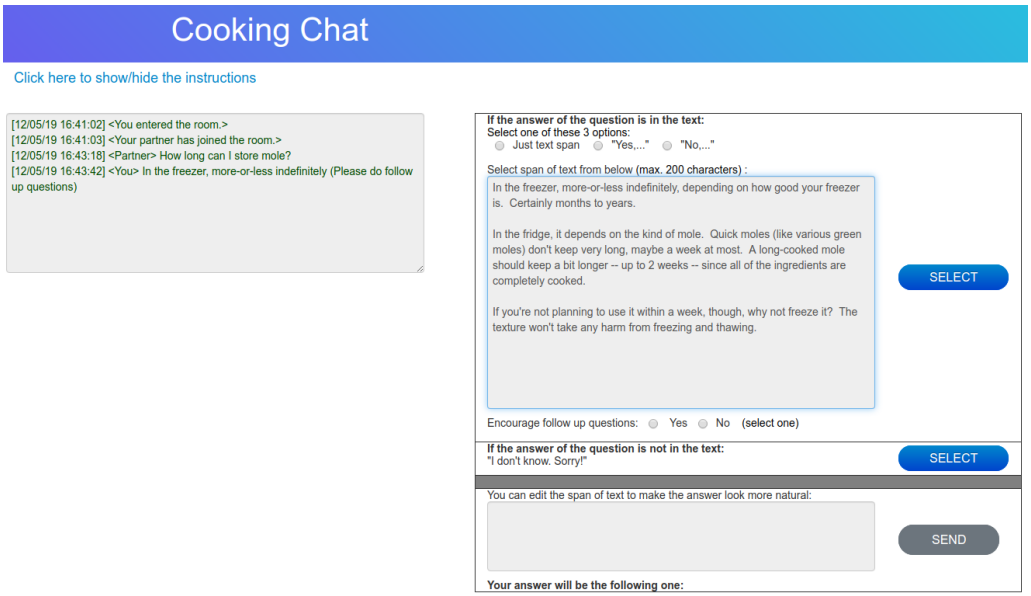


Figure 3: Dialogue collection interface for the expert.

Figure 4: Multiple answers collection interface.

Improving Conversational Question Answering Systems after Deployment using Feedback-Weighted Learning

Jon Ander Campos[1], **Kyunghyun Cho**[2], **Arantxa Otegi**[1],
Aitor Soroa[1], **Gorka Azkune**[1], **Eneko Agirre**[1]
[1]University of the Basque Country (UPV/EHU)
[2]New York University (NYU)
[1]{jonander.campos, arantza.otegi, a.soroa,
gorka.azkune, e.agirre}@ehu.eus, [2]kyunghyun.cho@nyu.edu

Abstract

The interaction of conversational systems with users poses an exciting opportunity for improving them after deployment, but little evidence has been provided of its feasibility. In most applications, users are not able to provide the correct answer to the system, but they are able to provide binary (correct, incorrect) feedback. In this paper we propose feedback-weighted learning based on importance sampling to improve upon an initial supervised system using binary user feedback. We perform simulated experiments on document classification (for development) and Conversational Question Answering datasets like QuAC and DoQA, where binary user feedback is derived from gold annotations. The results show that our method is able to improve over the initial supervised system, getting close to a fully-supervised system that has access to the same labeled examples in in-domain experiments (QuAC), and even matching in out-of-domain experiments (DoQA). Our work opens the prospect to exploit interactions with real users and improve conversational systems after deployment.

1 Introduction

In Conversational Question Answering (CQA) systems, the user makes a set of interrelated questions to the system, which extracts the answers from reference text (Choi et al., 2018). These systems are trained on datasets of human-human dialogues collected using Wizard-of-Oz techniques, where two crowdsourcers are paired at random to emulate the questioner and the answerer. Several projects have shown that it is possible to train effective systems using such datasets. For instance, QuAC includes question and answers about popular people in Wikipedia (Choi et al., 2018), and DoQA includes question-answer conversations on cooking, movies and travel FAQs (Campos et al., 2020). Building such datasets comes at a cost, which limits the widespread use of conversational systems built using supervised learning.

The fact that conversational systems interact naturally with users poses an exciting opportunity to improve them after deployment. Given enough training data, a company can deploy a basic conversational system, enough to be accepted and used by users. Once the system is deployed, the interaction with users and their feedback can be used to improve the system.

In this work we focus on the case where a CQA system trained off-line is deployed and receives explicit binary (correct, incorrect) feedback from users. An example of this task can be seen in Figure 1 where at a point in the conversation two different users give binary feedback to the system according to the correctness of the received answer. Assuming a large number of interactions, we can safely ignore examples for which no feedback is received. We propose feedback-weighted learning (FWL) based on importance sampling as the technique to improve the initial supervised system using only binary feedback from users.

In our experiments user feedback is simulated, and the correct/incorrect feedback is extracted from the gold standard. That is, if the system output matches the gold standard output then it is deemed correct, otherwise it is taken to be incorrect. In order to develop and test feedback-weighted learning we perform initial experiments on document classification. The results show that the model improved by the

Proceedings of the 28th International Conference on Computational Linguistics, pages 2561–2571
Barcelona, Spain (Online), December 8-13, 2020

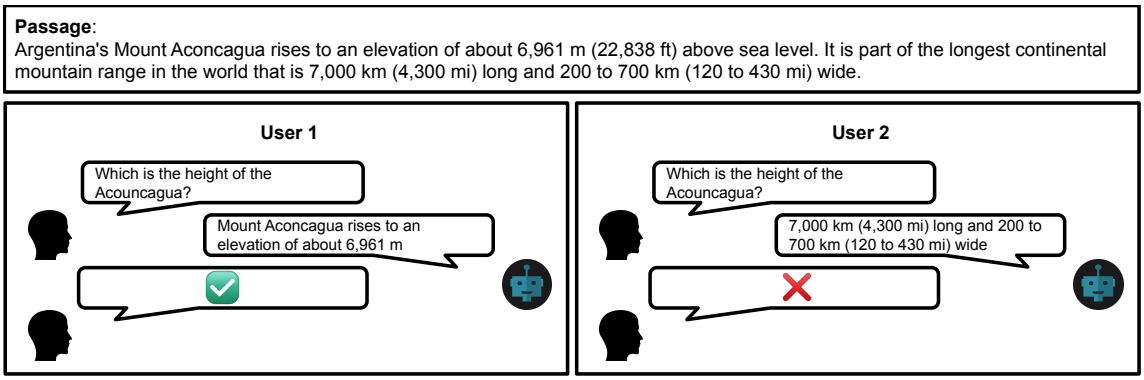


Figure 1: Example of the CQA task where at a point in the conversation the user 1 gives positive feedback to the system and user 2 gives a negative one due to the received incorrect answer.

proposed algorithm performs comparably to the fully supervised model that is fine-tuned with true labels rather than binary feedback. Those experiments are also used to check the impact of hyperparameters like the weight of the feedback and the balance between exploitation and exploration, which shows that our method is not particularly sensitive to the values of those hyperparameters.

Regarding CQA, we use the best hyperparameters from the earlier experiment on document classification, and conduct experiments using several domains in CQA including datasets like QuAC and DoQA. Our method always improves over the initial supervised system. In the in-domain experiments (QuAC) our method is close to the fully supervised model which is fine-tuned with true labels rather than binary feedback, and in the out-of-domain experiments (DoQA) our method matches it. The out-of-domain results are particularly exciting, as they are related to the case where a CQA system trained off-line in one domain could be deployed in another domain, letting the users improve it via their partial feedback by interacting with the system. Our experiments reveal that the proposed approach is robust to the choice of the system architecture, as we experimented with both multi-layer perceptron and pre-trained transformer.

The main contribution of our work is a novel method based on importance sampling, feedback-weighted learning, which improves the results of two widely used deep learning architectures using partial feedback only. Experimental results from document classification show that feedback-weighted learning improves over the initial supervised system, matching the performance of a fully supervised system which uses true labels. In-domain and out-of-domain CQA experiments show that the proposed method improves over the initial supervised system in all cases, matching a fully supervised system in out-of-domain experiments. This work opens the prospect to exploit interactions with real users and improve conversational systems after deployment. All the code and dataset splits are made publicly available [1].

2 Related Work on Conversational Question Answering

CQA research builds on reading comprehension. In reading comprehension the system has to answer questions about a certain passage of text in order to show that it understands the passage. There are two main methods: the *extractive* method, in which the answer is selected as a contiguous span in the reference passage, and the *abstractive* method, in which the answer text is generated. Many datasets (Rajpurkar et al., 2016; Rajpurkar et al., 2018; Dunn et al., 2017; Kočiskỳ et al., 2018; Trischler et al., 2017; Bajaj et al., 2016) and systems have been proposed to address this task, where the *extractive* scenario has drawn special attention (Wang and Jiang, 2017; Seo et al., 2017). Lately, with the incursion of large pre-trained language models as BERT (Devlin et al., 2019), XLNet (Yang et al., 2019) and their relatives, the state of the art has been dominated by systems that use the representations obtained with these pre-trained language models. The systems learn answer pointer networks that consist of two

[1] https://github.com/jjacampos/FeedbackWeightedLearning

classifiers, one for spotting the start token of the answer span and another for spotting the end token of the answer span. In reading comprehension, the questions are individual and isolated, that is, they do not have any dialogue structure.

Due to the increasing interest on modelling the conversational structure behind user questions, several CQA datasets where questions and answers are interrelated have been created following the Wizard-of-Oz technique. Among all the datasets we can highlight QuAC (Choi et al., 2018), CoQA (Reddy et al., 2019) and DoQA (Campos et al., 2020). While the first two datasets cover more formal domains as Wikipedia articles and literature, the latter covers different domains extracted from online forums as StackExchange. Contextual versions of the previously mentioned reading comprehension models have successfully modelled the conversational structure in those datasets (Qu et al., 2019b; Qu et al., 2019a; Ohsugi et al., 2019; Ju et al., 2019).

3 Importance Sampling for Learning After Deployment

In our learning after deployment scenario we start by training an initial S_0 system in an off-line and supervised way. This first system follows the traditional workflow where we have access to limited supervised training and development data. Then, we take the best performing system on the development data and deploy it to serve user queries. In this deployment phase, every time a user makes a query x, the system generates an answer y and the user gives binary feedback to it. Over time, the system generates different answers $y_{i1}, y_{i2}, ..., y_{in}$ and receives feedback for each item x_i . We assume a sufficient amount of user interactions, and as such we ignore any query-answer pair for which the user did not provide feedback. After the system has been deployed for a while, we collect for each question the answers provided by the system, and the respective user feedback.

We consider a CQA system implemented using two classifiers predicting the start and end tokens respectively. This allows us to consider each classifier independently and describe the process of learning after deployment for a single classifier. We propose to use feedback-weighted learning, which is based on self-normalized importance sampling, in order to generate the system answers.

3.1 Feedback-Weighted Learning

In this section, we describe a novel algorithm for updating a classifier trained off-line on-the-fly based on user feedback alone. We start by defining the true distribution $p^*(y|x)$ over C classes given an input x. This distribution is constructed to reflect binary user feedback $\{-\beta, \beta\}$:

$$p^*(y|x) \propto \begin{cases} \exp(\beta), & \text{if } y \text{ is correct} \\ \exp(-\beta), & \text{if } y \text{ is incorrect} \end{cases}$$

In words, the correctness of each class is reflected in the magnitude of the probability assigned to the class which is proportional to the user feedback. The hyperparameter β controls the weight of the feedback.

The goal of the proposed algorithm is to minimize the KL divergence from p^* to the classifier's predictive distribution $q(y|x;\theta)$ w.r.t. the parameters θ, where

$$\mathrm{KL}(p^* \| q) = -\sum_y p^*(y|x) \log q(y|x;\theta) + \mathcal{H}(p^*). \tag{1}$$

Exact minimization of this objective is however intractable due to the lack of access to the true distribution p^*. We can instead query the unnormalized p^* given the input x and a candidate class y.

We thus resort to self-normalized importance sampling with the following proposal distribution:

$$\hat{q}(y|x) = \lambda q(y|x;\theta) + (1-\lambda)\mathcal{U}(y), \tag{2}$$

where $\mathcal{U}(y)$ is a uniform distribution over y and smooths out the potentially peaky predictive distribution q. We can control this smoothness, which trades off exploration and exploitation, by controlling the mixing coefficient λ (Hoi et al., 2018).

With this proposal distribution, we derive the following objective function for feedback-weighted learning, starting from Eq. (1):

$$\begin{aligned}\mathrm{KL}(p^*\|q) - \underbrace{\mathcal{H}(p^*)}_{\text{const. w.r.t. } \theta} &= -\sum_y \hat{q}(y|x)\, \underbrace{\frac{p^*(y|x)}{\hat{q}(y|x)}}_{=w(y^k)} \log q(y|x;\theta) \\ &\approx -\frac{1}{K}\sum_{k=1}^{K} \frac{\omega(y^k)}{\sum_{k=1}^{K}\omega(y^k)} \log q(y^k|x;\theta), \end{aligned} \tag{3}$$

where K is the total number of user feedback received.

The importance weight $\omega(y^k)$ is computed as

$$\log \omega(y^k) = \underbrace{\beta \mathbb{1}(y^k = y^*)}_{=\text{feedback}} - \log \hat{q}(y|x), \tag{4}$$

where y^* is the (unknown) true class, and

$$\mathbb{1}(\alpha) = \begin{cases} 1, \text{ if } \alpha \text{ is true} \\ -1, \text{ if } \alpha \text{ is false} \end{cases}$$

In other words, the importance weight reflects the ratio between the user feedback and the model's confidence in each sampled prediction y^k. We hence call this algorithm *feedback-weighted learning*.

3.2 Related Work on Lifelong Learning

Continual or lifelong learning is defined as a system's ability to continually learn over time by accommodating new knowledge while keeping previously learned experiences (Parisi et al., 2019). Within this framework of lifelong learning, we particularly focus on building a system that adapts to changes in the data distribution after deployment (Agirre et al., 2019).

There have been efforts for learning actively from dialogue during deployment. The question answering (QA) setting was explored in Weston (2016) and Li et al. (2017), where they analyzed a variety of learning strategies for different dialogue tasks with diverse types of feedback. In these studies they also touch on *forward prediction*, which uses explicit user correction. This idea was later applied to chit-chat systems (Hancock et al., 2019). These works relied on users explicitly providing the correct answer. This strong assumption was relaxed in Weston (2016), where the user provides binary feedback on correct and incorrect answers in a synthetic question answering task (Weston et al., 2015). Our work also uses binary feedback and tests it in more realistic CQA datasets.

In a similar online setup to ours, Liu et al. (2018b) explored contextual multi-armed bandits for dialogue response selection using a customized version of Thompson sampling. In this work they use the Ubuntu Dialogue Corpus (Lowe et al., 2015) for user simulation. In the case of task-oriented dialogue systems, Liu et al. (2018a) propose a hybrid learning method with supervised pre-training and further improvement using human teaching and feedback. For the human teaching case they use imitation learning with explicit corrections done by an expert. After that, they resort to reinforcement learning for further improvement thanks to long term rewards defined by task completion.

4 Experiments

In this section we present the experiments with feedback-weighted learning (FWL). In the experiments we first build a supervised system (S_0), and then we simulate a deployment phase by letting S_0 answer user queries and receiving their feedback. User feedback is derived from a manually annotated deployment set, which is obtained by splitting the training set. We refer to the set used for training S_0 as a *training set* and the other partition of the original training set as a *deployment set* in the rest of the paper.

We consider the following systems and baselines:

- S_0: the original supervised system trained on the training dataset only. We consider this system a baseline.

- S_0 + *FWL*: S_0 is fine-tuned with FWL using examples and partial feedback from the deployment set.

- S_0 + *supervised*: we first train S_0 as above, and then continue its training using examples from the deployment set using the true labels instead of binary feedback. This is thus a fine-tuned system that has full access to the true data.

- *Fully supervised*: a supervised system trained from scratch using the union of the training and deployment sets.

Although our main objective is to develop a lifelong learning system for CQA, we also perform experiments on document classification, as a way to assess the robustness of the proposed method when applied to different neural architectures and tasks. Moreover, these experiments are used to develop the system and check the impact of hyperparameters, so that the best hyperparameters from document classification are used in the CQA experiments.

4.1 Document Classification

The model for document classification is a simple multi layer perceptron (MLP) with a single hidden layer. The input to the MLP is a document vector, calculated as the average of the GloVE vectors (Pennington et al., 2014) of all the words in the document. The dimension of the embeddings is set to 300, and the hidden layer has 200 hidden units.

Experiments are performed on the DBPedia Classes dataset,[2] which contains hierarchical categories of $342,748$ Wikipedia articles. Each article is categorized at three levels into 9, 70 and 219 categories respectively. We use the latter setting with 219 classes in our experiments. The dataset comes with a standard train, development and test splits. We kept the development and test sets untouched, but we split the training part further, creating a training set and a deployment set with the 10% and 90% of the original training examples, respectively. These percentages are motivated on real scenarios where the initial amount of training data is usually limited and expensive to obtain, but during deployment it could be easier to collect more data in a cheaper way. In the deployment phase we consider the feedback to be positive when the class assigned by the system is the same as the gold class in the deployment set, and negative otherwise.

Regarding the experimental setting, the S_0 system is built on the train split using cross entropy loss. For the S_0 + *FWL* system we perform hyperparameter exploration of $\lambda \in [0.5, 1.0]$ and $\beta \in [1, 85]$ using Bayesian optimization (Snoek et al., 2012). The hyperparameter values that performed best in the original development set after one epoch are selected, which corresponds to $\lambda = 0.97$ and $\beta = 76$. We sample class predictions 3 times for each example, based on our preliminary experiments, and train S_0 + *FWL* a maximum of 50 epochs. Given N the amount of training examples and K the amount of samples, in this article we will use *epoch* to mean $N \times K$ feedback requests. See Section 5 for a further discussion on sample efficiency in FWL.

Table 1 shows that the simple MLP architecture performs well on this task, even when only the 10% of training examples are used. Still, S_0 + *FWL* is able to improve the performance of S_0 by 5 points, and it is close to both supervised systems. These results validate the effectiveness of FWL as a way of improving an initial supervised system using binary feedback only.

4.2 Conversational Question Answering

In the CQA experiments we fine-tune a pretrained BERT (Devlin et al., 2019) for QA. Given a query and a passage that contains the answer, the pretrained BERT is fine-tuned to predict the start and end indexes of the answer span. This approach has shown strong performance on QA datasets such as SQuAD

[2] https://www.kaggle.com/danofer/dbpedia-classes

Systems	F1
S_0	86.51
S_0 + FWL	91.59 (+5.0)
S_0 + supervised	91.89 (+5.3)
Fully supervised	92.04 (+5.5)

Table 1: Results as F1 on document classification. Number in parenthesis for difference with respect to S_0. FWL continues learning over S_0 using only binary feedback, and the result is close to the supervised systems.

Systems	no history	dialogue history
S_0	46.76	49.03
S_0 + FWL	49.33 (+2.6)	53.07 (+4.0)
S_0 + supervised	53.66 (+6.9)	55.10 (+6.1)
Fully supervised	54.50 (+7.7)	55.40 (+6.5)

Table 2: Results of in-domain experiments using QuAC dataset both for training and deployment, with and without dialogue history. F1 accuracy results on QuAC development split. Number in parenthesis for difference with respect to S_0. FWL is able to improve over S_0 which validates its usefulness in CQA.

(Rajpurkar et al., 2016). In our experiments we use the base uncased model of BERT with the maximum context size of 384 and a batch size of 12, using default values for the rest of the hyperparameters.

We experiment with the following settings:

- In-domain vs. out-of-domain. We experiment with two different scenarios, based on the mismatch between training and deployment distributions. In the first scenario the domain is the same for both training and deployment phases, whereas in the out-of-domain scenario the domains differ.
- Without vs. with dialogue history. In order to take into account the multi-turn feature of a dialogue, we prepend the previous question and its corresponding answer to the input. Following usual practice (Qu et al., 2019a), we consider only the previous interaction (one questions and one answer).

In the in-domain experiments we use QuAC (Choi et al., 2018) for both building the initial S_0 system and during the deployment phase. QuAC is a conversational dataset extracted from the Wikipedia using the Wizard of Oz method and crowdsourcing. In the out-of-domain scenario QuAC is used for building S_0, but the deployment phase is done with DoQA (Campos et al., 2020), which is a conversational dataset based on FAQs and contains dialogues from three different domains (cooking, travel and movies).

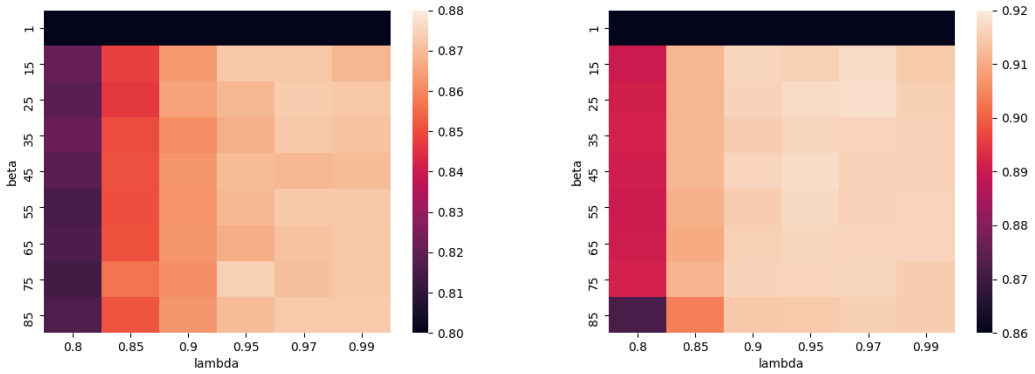
Similarly to document classification, we split the original training parts of QuAC into training and deployment splits containing 10% and 90% of the training dialogues, respectively. We consider the feedback to be positive whenever the answer span predicted by the system matches the gold span exactly, and negative otherwise. Because the QuAC test split is unavailable, we report results in the development split.

With respect to the system settings used for the experiments, we set the λ and β hyperparameters of the S_0 system based on their best values from document classification ($\lambda = 0.97$ and $\beta = 76$). Given that the CQA system contains two classifiers and the number of classes is often larger than in document classification task, we use a larger number of samples, 50 in this task.

Table 2 shows the results on the in-domain experiments on the QuAC dataset. For each system we report the results after 3 epochs following Qu et al. (2019a). The results follow the trend observed in the document classification setting. Applying FWL after S_0 improves the results by 2.6 and 4 points, which confirms that FWL is a valid technique to continue fine-tuning a CQA system after deployment. Using dialogue history improves the results of all systems by almost 3 points, stressing the importance of modeling history on CQA systems. However, the main conclusions remain unchanged. S_0 + *FWL* still outperforms S_0 using only binary feedback, and is close to the supervised systems.

Systems	Cooking	Movies	Travel
S_0	39.79	40.89	35.64
S_0 + FWL	49.66 (+9.9)	47.28 (+6.4)	47.19(+11.6)
S_0 + supervised	50.63 (+10.8)	46.79 (+5.9)	47.12(+11.5)
Fully supervised	50.33 (+10.5)	45.56 (+4.7)	46.10(+10.5)

Table 3: Results of out-of-domain experiments (with history modeling) using QuAC for training and DoQA during deployment. F1 accuracy results on DoQA test split on cooking, movies and travel domains. Number in parenthesis for difference with respect to S_0. FWL improves the results of S_0 and matches supervised results in two domains.



(a) F1 scores obtained after one epoch and using 3 samples

(b) F1 scores obtained after 50 epochs and using 3 samples

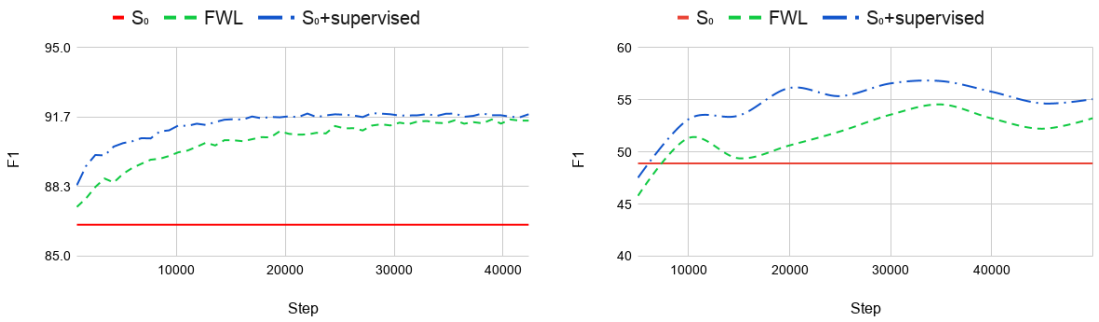
Figure 2: Hyperparameter analysis using heatmaps on document classification showing the obtained F1 scores (lighter is better) in the development split. Similar performance is obtained with different hyperparameter pairs, showing the robustness of the method.

Table 3 shows the results when S_0 is trained on QuAC, and the user feedback is simulated using examples from DoQA. In these experiments we perform model selection on the development split of DoQA (which corresponds to the cooking domain) and report the results on the test datasets comprised of the cooking, travel and movies. We report only experiments using dialogue history, as this setting is more realistic for a CQA system. S_0 + *FWL* outperforms S_0 across all the domains. S_0 + *FWL* furthermore matches the S_0 + *supervised* system in the movies and travel domains, although it fails to do so in the cooking domain. The fully supervised system performs worse than S_0+ *supervised* on this dataset, which we conjecture is due to the fact that QuAC contains more training examples than DoQA, with a ratio of approximately 3 to 1. This may cause the fully supervised system to be more biased towards QuAC, and thus yields worse results in DoQA. Note that in the S_0 + *supervised* system QuAC examples are used to train S_0 only, which is then fine-tuned with DoQA examples, and obtains better results overall. All in all, these results suggest that the FWL approach is robust when there is a domain shift between the training and test datasets.

5 Discussion

As shown by the experiments in document classification and CQA we are able to improve an initial supervised S_0 system just by using binary feedback obtained by simulating the users. In this section we perform a further analysis on several aspacts of the method.

Hyperparameters. In order to show the robustness of FWL we perform several experiments in the document classification task with different values for the main hyperparameters of the method, λ and β (cf. Section 3.1). The analysis shows that when using values larger than 1 for β, FWL performs similarly



(a) Document classification

(b) CQA, in-domain (QuAC dataset), with dialogue history

Figure 3: Learning curves for the document classification and CQA tasks where FWL is compared to supervised learning. As the number of steps increase FWL gets closer to S_0 + *supervised*.

well for all lambdas greater than 0.8 (see Figures 2a and 2b). The behavior of λ in the same Figures 2a and 2b reveals that large values of λ yields best results for all beta values. In any case, the similar performance obtained with different hyperparameter combinations shows that our method is robust and not specially sensitive to small variations in the hyperparameters.

Learning dynamics. From the learning curves in Figures 3a and 3b we see how the behavior is similar in both document classification and CQA learning tasks. In both cases the supervised systems converge faster than the FWL systems, but as the steps go on the F1 scores in the development set also converge. It is of special interest the point where FWL improves over S_0. In the document classification task FWL improves over S_0 in the first steps, and by the end of the first iteration, which comprises circa 850 steps, it already outperforms S_0. In CQA FWL needs more steps but the improvement over S_0 also happens at the beginning of the training process.

Sampling vs. supervised learning. Since we treat epochs in FWL as in supervised learning, we sample new answers for each new epoch. For example, in the document classification case we end up taking 150 samples (50 epochs with 3 samples per epoch) for a total of 219 classes. It can be argued that a dummy sampling technique covering all classes is equivalent to having the true label, and would be similar to our method in terms of sampling efficiency. However, when deploying a $S0$ system in a realistic scenario, the dummy sampling strategy would return low probability responses and could severely hamper user engagement. In contrast, our sampling method tends to return high probability answers, making it more user-friendly. In any case, each time the loss gradient is computed, FWL has information of only 3 samples, unlike supervised learning where all classes are considered. Besides, 3 samples per example (one epoch) are enough for FWL to improve over S_0 (see Figure 2b), although the best results are obtained after 50 epochs.

Assumptions and limitations. We discuss a few assumptions we made in designing the proposed FWL. In all our experiments we simulate user feedback using supervised data, and thus the feedback is always accurate and explicit. We therefore do not consider the case where the user is unsure about the response it gave to the system, which would cause a noisy feedback that can harm the performance of the system. Moreover, as we need more than one sample for each question we would need different users making the same questions if we were to compare our method with real use-cases. Analyzing the impact of these issues and possible solutions to them is kept as an open research question for future analysis.

6 Conclusion and Future Work

In this work we propose feedback-weighted learning that allows a supervised classifier to effectively adapt itself after deployment from partial user feedback. The experiments show that our technique is successful, in that it improves over the initial supervised system. More specifically, in document classi-

fication experiments, it matches an off-line supervised system trained with all the true labels, although it has only access to the binary feedback. More importantly, the experiments in two widely used CQA datasets, QuAC and DoQA, confirm that it is feasible to improve a CQA system after deployment. In the DoQA experiments, the CQA system is trained off-line in one domain (Wikipedia) and then deployed in other domains, letting the users improve it via their partial feedback by interacting with the system. In this setting, the performance of our model also matches that of the fully supervised model which is fine-tuned with true labels rather than binary feedback. Moreover, feedback-weighted learning is shown to be effective in two deep learning architectures, including a multi-layer feed forward network and a high-performing pre-trained transformer fine-tuned in the task.

This work uses simulated feedback derived from gold standard labels. In the future we plan to modify feedback-weighted learning to cope with noisy feedback, as well as modifying it to work with fewer samples per query.

Acknowledgments

This research was partially supported by a Google Faculty Award, EU ERA-Net CHIST-ERA LIHLITH funded by the Agencia Estatal de Investigación (AEI, Spain) project PCIN-2017-118, the Basque Government excellence research group (IT1343-19) and the NVIDIA GPU grant program. Jon Ander Campos enjoys a doctoral grant from the Spanish MECD. Kyunghyun Cho was partly supported by Samsung Advanced Institute of Technology (Next Generation Deep Learning: from pattern recognition to AI) and Samsung Electronics (Improving Deep Learning using Latent Structure), and thanks CIFAR, eBay, NVIDIA and NAVER for their support.

References

Eneko Agirre, Anders Jonsson, and Anthony Larcher. 2019. Framing Lifelong Learning as Autonomous Deployment: Tune Once Live Forever. In *International Workshop on Spoken Dialogue Systems Technology*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. MS MARCO: A Human-Generated MAchine Reading COmprehension Dataset. *arXiv preprint arXiv:1611.09268*.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - Accessing Domain-Specific FAQs via Conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314, Online, July. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. *arXiv preprint arXiv:1704.05179*.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from Dialogue after Deployment: Feed Yourself, Chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684.

Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. 2018. Online learning: A Comprehensive Survey. *arXiv preprint arXiv:1802.02871*.

Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on Conversational Question Answering. *arXiv preprint arXiv:1909.10772*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2017. Dialogue Learning With Human-in-the-Loop. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018a. Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems. In *Proceedings of NAACL-HLT*, pages 2060–2069.

Bing Liu, Tong Yu, Ian Lane, and Ole J Mengshoel. 2018b. Customized Nonlinear Bandits for Online Response Selection in Neural Conversation Models. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294.

Yasuhito Ohsugi, Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2019. A Simple but Effective Method to Incorporate Multi-turn Context with BERT for Conversational Machine Comprehension. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 11–17.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual Lifelong Learning with Neural Networks: A review. *Neural Networks*, 113:54–71.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019a. BERT with history answer embedding for conversational question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1133–1136.

Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W Bruce Croft, and Mohit Iyyer. 2019b. Attentive History Selection for Conversational Question Answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1391–1400.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional Attention Flow for Machine Comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2951–2959. Curran Associates, Inc.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

Shuohang Wang and Jing Jiang. 2017. Machine Comprehension Using Match-LSTM and Answer Pointer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards AI complete Question Answering: A Set of Prerequisite Toy Tasks. *arXiv preprint arXiv:1502.05698*.

Jason E Weston. 2016. Dialog-based Language Learning. In *Advances in Neural Information Processing Systems*, pages 829–837.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

A.2 DoQA conversation examples

In this appendix we present conversation examples from the DoQA dataset.

Dialogue 1

Question: Smoked a chicken and the skin came out like boot leather! What can I do?

Answer : protect the skin during the long slow smoking process. You may want to pull the cheesecloth off about 30 minutes before you are done.(Please instead of following up, try additional questions)

Question: Is 4 hours too long to keep in the smoker?

Answer : I don't know. Sorry!

Question: What about using a little apple juice and vegetable oil on the skin?

Answer : yes you can(Please instead of following up, try additional questions)

Question: What is a good temperature for the smoker?

Answer : 400 degrees(Please instead of following up, try additional questions)

Question: Do you have any other tips to prevent this from happening again?

Dialogue 2

Question: Does poking steaks and other meats cause flavor loss?

Answer : Yes, "This is totally fine. Meat gets poked and prodded all the time, whether to test it, to pick it up with tongs, or sticking a fork in it to cut. Meat just is not that fragile. (Please do follow up questions)

Question: Is there a way I can alter the flavor on the meat?

Answer : I don't know. Sorry!

Question: How should I reduce the "dead taste"?

Answer : Yes, Things like smashing with a tenderizer (Please do follow up questions)

Question: Great, I will give that a try on my meat. A friend said that I'm ruining the steak and that it causes flavor and juice losses. Is this the case?

Answer : Meat just is not that fragile.(Please do follow up questions)

Question: Okay, how should I place the meat on the grill?

Answer : I don't know. Sorry!

Question: Okay thanks!

Dialogue 3

Question: Ginger starting to sprout - can I still use it?

Answer : Yes, I find that when sprouted, the ginger just loses a bit of flavour, that's all. (Please do follow up questions)

Question: Do I have to prepare it in a different way than non sprouted ginger?

Answer : I don't know. Sorry!

Question: Does it have a different taste than non sprouted ginger?

Answer : Yes, the ginger just loses a bit of flavour, that's all. If you let it grow for a very long time and it becomes all shriveled, then you are talking about a flavourless piece of ginger. (Please do follow up questions)

Question: What about the sprouts, are they edible?

Answer : I don't know. Sorry!

Question: Can I plant sprouted ginger to make more ginger roots?

Answer : I don't know. Sorry!

Dialogue 4

Question: In general, can I substitute egg species for either egg-only or egg-centric dishes?

Answer : There may be some issues with substituting in dishes due to varying fat content of the yolks and such...(Please do follow up questions)

Question: Do you think that the smaller eggs like quail eggs would have more flavor for a quiche?

Answer : I don't know. Sorry!

Question: I think that duck eggs would make a great addition to chili rellenos because of the high fat content. Do you think that would be something great to have for breakfast?

Dialogue 5

Question: What is an arepa?

Answer : White pre-cooked corn flour, water, pinch of salt and cooked in the pan(Please instead of following up, try additional questions)

Question: Can you use anything besides water to cook them?

Answer : Yes, cooked in the pan and in the oven and decide for yourself which you prefer (Please do follow up questions)

Question: Is it better to cook them in the oven or in the pan?

Answer : I don't know. Sorry!

Dialogue 6

Question: Fluffy texture in a Spanish tortilla

Answer : "If you want it fluffy, you'll need air(Please do follow up questions)

Question: How can I add air to it?

Answer : The easiest way to accomplish this is to crack the eggs in a bowl, whip them (with a whisk or a fork) for a couple of minutes(Please do follow up questions)

Question: Do I add it to a hot pan?

Answer : I don't know. Sorry!

Question: When do I put it under the grill?

Answer : I don't know. Sorry!

Question: Do I add spices?

Answer : I don't know. Sorry!

Dialogue 7

Question: Does poking steaks and other meats cause flavor loss?

Answer : Yes, "This is totally fine. Meat gets poked and prodded all the time, whether to test it, to pick it up with tongs, or sticking a fork in it to cut. Meat just is not that fragile. (Please do follow up questions)

Question: Is there anything I could do wrong to give the meat a 'dead taste'?

Answer : I don't know. Sorry!

Question: How much moisture will a steak lose whilst cooking?

Answer : Yes, Things like smashing with a tenderizer are much more violent and do affect the texture (that's kind of the point), so that's not a great comparison. But this is a really minor thing (Please do follow up questions)

Question: Would you not recommend using a tenderizer?

Answer : No, This is totally fine. (Please do follow up questions)

Question: how long should meat be tenderized for ?

Answer : I don't know. Sorry!

Dialogue 8

Question: Does poking steaks and other meats cause flavor loss?

Answer : This is totally fine. Meat gets poked and prodded all the time, whether to test it, to pick it up with tongs, or sticking a fork in it to cut. Meat just is not that fragile.(Please instead of following up, try additional questions)

Question: Would you recommend using a meat tenderizer?

Answer : Things like smashing with a tenderizer are much more violent and do affect the texture (that's kind of the point), so that's not a great comparison. But this is a really minor thing(Please instead of following up, try additional questions)

Question: Is it possible to bruise the steak using a tenderiser?

Answer : Things like smashing with a tenderizer are much more violent and do affect the texture (that's kind of the point), so that's not a great comparison. But this is a really minor thing.(Please do follow up questions)

Question: How does tenderising affect the flavour?

Answer : I don't know. Sorry!

Question: Would you suggest using a flavour enhancer?

Answer : I don't know. Sorry!

Dialogue 9

Question: How do I know when a chicken breast has cooked through?

Answer : Yes, A thermometer is the only way to be sure. (Please do follow up questions)

Question: Are there other ways to tell? In the event I did not have a thermometer for some reason.

Answer : Yes, you should able to learn the average cooking time, and outwards cues of color and texture that match the right internal temperature. (Please do follow up questions)

Question: What are some of the outward cues?

Answer : Internally, the meat should look opaque and white.(Please do follow up questions)

Question: Okay, how about externally?

Answer : I don't know. Sorry!

Question: What temperature should the meat be cooked at in order to cook all the way through?

Answer : I don't know. Sorry!

Question: Is cutting the meat open a good way to make sure it is cooked

through?

Answer : No, For methods with a consistent level of heat (stove, oven), you should able to learn the average cooking time, and outwards cues of color and texture that match the right internal temperature. (Please do follow up questions)

Question: What should the internal temperature be exactly?

Answer : I don't know. Sorry!

Dialogue 10

Question: I smoked a chicken and the skin came out like boot leather, can you help me figure out what went wrong?

Answer : Yes, "Take cheesecloth and soak it in melted butter and drape it over the bird before you put it in the smoker. This will protect the skin during the long slow smoking proces (Please do follow up questions)

Question: Okay, thanks. I had the temperature at 225-250 for 4 hours. Are both of those find when I try this again in the future?

Answer : I don't know. Sorry!

Question: If I do not have a cheesecloth is there something else I could use?

Answer : melted butter (Please instead of following up, try additional questions)

Question: What it I do not have melted butter to soak the cheesecloth in? Is there a substitute?

Answer : I don't know. Sorry!

Question: How much melted butter should I use?

Answer : I don't know. Sorry!

Dialogue 11

Question: Filtered or non-filtered soymilk maker? Filtered or non-filtered soymilk maker, which one would you suggest buying?

Answer : Yes, I have an old SoyaQuick (mine has a filter, newer models don't), and I think it was largely a cleaning concern. (Please do follow up questions)

Question: Do you most people will prefer using a newer version of the soyamilk maker?

Answer : I don't know. Sorry!

Dialogue 12

Question: Is it possible to cook a meatloaf using clear glass Pyrex containers?

Answer : Yes, "I see no reason you couldn't use that Pyrex set for a meatloaf - I've used glass casseroles for meatloaf before (so glass in general is no problem), and that set says the bowls are oven safe. (Please do follow up questions)

Question: Great thanks! I've got a pyrex set...are they generally microwave safe?

Answer : I don't know. Sorry!

Question: No worries...do you know if they are oven safe?

Answer : Yes, the bowls are oven safe. (Please instead of following up, try additional questions)

Question: Ok cool. What was the result when you cooked meatloaf using pyrex...big hit, or total disaster?

Answer : Yes, I'd recommend wrapping it in aluminum foil to help it keep its shape. (Please instead of following up, try additional questions)

Question: Oh is that what you did?

Answer : Yes, As for the cookie sheet method, I would be afraid of it falling apart as you described, but if you were to go that route, I'd recommend wrapping it in aluminum foil to help it keep its shape. (Please do follow up questions)

Question: Cool...How large was the meatloaf you cooked?

Dialogue 13

Question: Ginger starting to sprout - can I still use it?

Answer : Yes, From a culinary perspective, I find that when sprouted, the ginger just loses a bit of flavour, that's all. (Please do follow up questions)

Question: Can I use it just like regular ginger, or is there something different I should do with it?

Answer : I don't know. Sorry!

Question: Does it require special preparation?

Answer : I don't know. Sorry!

Dialogue 14

Question: Is gelatin vegetarian?

Answer : Gelatin comes from a dead animal (unless they start harvesting it

with arthroscopic probes :), so it is not a vegetarian ingredient. (Please do follow up questions)

Question: Is there a vegetarian alternative?

Answer : Yes, Yes, There are many other hydrocolloids, such as agar, that can be used to produce similar textures if needed. (Please instead of following up, try additional questions)

Question: So can you tell me what constitutes an ingredient as "vegetarian" or "vegan?"

Answer : I don't know. Sorry!

Question: That's ok. I was wondering if you had any recipes that include vegetarian gelatin?

Answer : I don't know. Sorry!

Dialogue 15

Question: What features should I look for when buying an espresso maker?

Answer : since it is the high-pressure components that are key to how they work, it is the quality and durability of these that tends to set the price point.(Please do follow up questions)

Question: Would a $30 machine have high pressure components like you mentioned?

Answer : I don't know. Sorry!

Question: Do you know what the differences between a $30 machine and a $500 machine are?

Answer : Yes, you generally get what you pay for with espresso machines. The more expensive domestic ones usually really do last much longer. (Please do follow up questions)

Question: What specific features should I look for?

Answer : My advice is to visit a couple of retailers that have staff dedicated to selling espresso machines.(Please do follow up questions)

Question: Will they be able to answer questions about different priced espresso machines?

Answer : Yes, they will also recognise when someone needs a domestic unit designed for daily use as opposed to one just for special occasions. (Please do follow up questions)

Dialogue 16

Question: Can I use garlic leaf for cooking?

Answer : Yes, When we have had garlic in our garden I have used the garlic leaves (Please do follow up questions)

Question: Can I dry the garlic leaves?

Answer : Regarding drying them, I have never tried it. Off the top of my head I can't think of any reason not to dry them for later use(Please do follow up questions)

Question: Would I be able to use the dried leaves like I use other herbs in cooking?

Answer : I don't know. Sorry!

Question: Do the leaves taste just like garlic?

Answer : They do have a garlicky flavor but are milder than garlic cloves(Please do follow up questions)

Question: Are they safe to eat?

Answer : Yes, tend to use them more as I would chives or garlic chives as in addition to having the milder flavor than the cloves they make for a quite nice presentation (Please do follow up questions)

Dialogue 17

Question: How long can you keep chocolate, and what is the best way to store it?

Answer : Regardless of type, all chocolate should be stored in a cool and low humidity (dry) place away from direct sunlight(Please do follow up questions)

Question: How long does chocolate last before losing flavor?

Answer : Dark chocolate will last for years. Milk and white chocolate will last for a much shorter time (a few months), because of their milk content.(Please do follow up questions)

Question: Once it gets that white stuff on the outside, is it done?

Answer : No, This kind of chocolate is still suitable for any application where the chocolate will be fully melted (most baking) (Please do follow up questions)

Question: What's the best way to store it for as long as possible?

Answer : all chocolate should be stored in a cool and low humidity (dry) place away from direct sunlight. It would be best to seal it in an air-tight container(Please do follow up questions)

Question: How long can I keep milk chocolate if properly stored?

Answer : Dark chocolate will last for years. Milk and white chocolate will last for a much shorter time (a few months), because of their milk content.(Please instead of following up, try additional questions)

Question: How do I know if chocolate has gone bad?

Answer : Improperly stored chocolate will develop bloom, which shows as a white or grey streaking or spotting on the surface.(Please do follow up questions)

Question: Will I get sick if I eat chocolate that has developed bloom?

Answer : No, This kind of chocolate is still suitable for any application where the chocolate will be fully melted (Please do follow up questions)

Dialogue 18

Question: What is the best way to store and manage tahini?

Answer : just spend some time and elbow grease to mix it back together again.(Please do follow up questions)

Question: What do I do when it separates?

Answer : When that happens, just spend some time and elbow grease to mix it back together again.(Please instead of following up, try additional questions)

Question: How do I recover it?

Answer : I don't know. Sorry!

Question: What are alternatives besides food processors?

Answer : I don't know. Sorry!

Question: What is the best way to store longterm?

Answer : if the tahini was stored in the fridge, it might take longer because everything will be harde(Please do follow up questions)

Question: Is there a better way?

Answer : I don't know. Sorry!

Dialogue 19

Question: How to cook good "arepas"?

Answer : I strongly recommend you experiment with the different flours, milk or water or half and half,(Please do follow up questions)

Question: If you wanted to sub corn flour would you just do it measure for measure? Sorry I meat white wheat flour

Answer : cookery is a living and evolving subject and very much a matter of personal taste. (Please do follow up questions)

Question: do you use water or milk?

Answer : water,(Please do follow up questions)

Question: and which is better oven or pan?

Answer : cooked in the pan is the traditional way(Please instead of following up, try additional questions)

Question: are these a sweet or savory food?

Answer : White pre-cooked corn flour, water, pinch of salt(Please do follow up questions)

Dialogue 20

Question: How good a substitute is callaloo for spinach?

Answer : d amaranth, so apologies if that is distinct from the type you have access to. As I recall, spinach is a bit sweeter and the leaves are a bit softer so they break down more readily(Please do follow up questions)

Question: Have you used both spinach and amaranth in recipes?

Answer : I've only had red amaranth,(Please do follow up questions)

Question: Do you have any recipes that you would recommend for red amaranth?

Answer : I don't know. Sorry!

Question: Would you recommend using amaranth as a replacement if someone cannot get spinach?

Answer : Yes, I've yet to find a recipe so touchy that one leafy green can't be substituted for another. (Please do follow up questions)

Question: Do you have a favorite leafy green recipe?

Answer : I don't know. Sorry!

Question: OK, thank you

Dialogue 21

Question: Why does this work? (defrosting steak)

Answer : he reason is, solids and liquids transfer heat better than gasses do. (Please do follow up questions)

Question: would you recommend this over a microwave?

Answer : Yes, Sandwich the steak between two pots, one of which has a large mass of warm water in it; the heat from the water will flow into the meat. (Please do follow up questions)

Question: How long would you expect it to take for a medium sized steak ?

Answer : I don't know. Sorry!

Question: Is there any damage to the steak with this method?

Answer : No, The reason is, solids and liquids transfer heat better than gasses do. (Please do follow up questions)

Question: Does the steak start to brown slightly?

Answer : I don't know. Sorry!

Dialogue 22

Question: Can I use garlic leaf for cooking?

Answer : Yes, They do have a garlicky flavor but are milder than garlic cloves. (Please do follow up questions)

Question: Is it safe to use?

Answer : I don't know. Sorry!

Question: Should I dry them?

Answer : Regarding drying them, I have never tried it. Off the top of my head I can't think of any reason not to dry them for later use but there may be issues that I just don't know about.(Please instead of following up, try additional questions)

Question: Do you know what the measurements would be compared to regular garlic?

Answer : I don't know. Sorry!

(04:47:10) **Question**: What types of food is garlic used for?

Answer : I don't know. Sorry!

Dialogue 23

Question: Why can't this ice cream scoop go in the dishwasher?

Answer : Yes, "I've accidentally run my scoop, a Zeroll with conductive fluid inside the handle, through the dishwasher. (Please do follow up questions)

Question: Will this ruin the scoop part of the ice cream scoop?

Answer : Yes, is that the fluid is meant to work at normal body temperature and when it gets too hot, like in a dishwasher, it solidifies (Please do follow up questions)

Question: what is the conductive fluid made of?

Answer : I don't know. Sorry!

Dialogue 24

Question: Why can't this ice cream scoop go in the dishwasher?

Answer : I believe what happened to mine (and what's happened to yours) is that the fluid is meant to work at normal body temperature and when it gets too hot, like in a dishwasher, it solidifies.(Please instead of following up, try additional questions)

Question: Does it do something to the metal?

Answer : I believe what happened to mine (and what's happened to yours) is that the fluid is meant to work at normal body temperature and when it gets too hot, like in a dishwasher, it solidifies.(Please instead of following up, try additional questions)

Question: Can I buy one that can go into the dishwasher?

Answer : I don't know. Sorry!

Question: If I was making india curry and don't have yogurt is there something I can substitute

Answer : I don't know. Sorry!

Question: If I wasn't going to used a ice cream spoon what else can I used to replace it

Answer : I don't know. Sorry!