Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 436–442
Marseille, 11–16 May 2020

Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque

Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa, Eneko Agirre
University of the Basque Country UPV/EHU
{arantza.otegi, jonander.campos, a.soroa, e.agirre}@ehu.eus
aagirre014@ikasle.ehu.eus

Abstract

Conversational Question Answering (CQA) systems meet user information needs by having conversations with them, where answers to the questions are retrieved from text. There exist a variety of datasets for English, with tens of thousands of training examples, and pre-trained language models have allowed to obtain impressive results. The goal of our research is to test the performance of CQA systems under low-resource conditions which are common for most non-English languages: small amounts of native annotations and other limitations linked to low resource languages, like lack of crowdworkers or smaller wikipedias. We focus on the Basque language, and present the first non-English CQA dataset and results. Our experiments show that it is possible to obtain good results with low amounts of native data thanks to cross-lingual transfer, with quality comparable to those obtained for English. We also discovered that dialogue history models are not directly transferable to another language, calling for further research. The dataset is publicly available.

Keywords: Conversational Question Answering, pre-trained language models, BERT, cross-lingual transfer, Basque

1. Introduction

Conversational Question Answering (CQA) systems meet user information needs by having conversations with them. Users pose initial queries in free form text, and the systems usually answer the queries by returning relevant excerpts extracted from a reference passage. The answers returned by the system invite users to pose follow up questions, which are again answered, therefore creating a conversation between the system and the user.

The field has received much attention in the last years, and there exist nowadays a variety of datasets for the task (Rajpurkar et al., 2016; Trischler et al., 2017; Nguyen et al., 2016; Kočiský et al., 2018; Dunn et al., 2017; Choi et al., 2018). Some of the datasets are very large. For instance, QuAC (Choi et al., 2018) contains thousands of dialogues and tens of thousands of question answering turns, which have been collected using wizard-of-oz techniques with paid crowdworkers. The dataset is built on top of Wikipedia sections about popular people and organizations. The high results of current systems are encouraging, and seem to show that the technology is ready for industrial adoption. Unfortunately, all current datasets are in English, and data gathering is extremely expensive, which means that analogous CQA systems in other languages require high annotation budgets. In smaller language communities, there is the added problem of not having a critical mass of crowdworkers in the target language, which makes the collection of the question and answer conversations more difficult. To make matters more challenging, the size and amount of Wikipedia articles is lower, making it more difficult to automatically select popular topics with enough text. Although we focus on Basque in this paper, note that this harder conditions are not an issue only for Basque, but for most low resource languages as well.

In this paper, we test whether there is a real need to gather such large amounts of QA conversations in other languages, and to evaluate how far can we go with English training data and none or small amounts of native training data. We present ElkarHizketak, a small dataset of CQA interactions, analogous to QuAC but for Basque. Due to the lack of Basque speakers in crowdsourcing platforms, we used social media to recruit volunteers. The resulting dataset contains close to 400 dialogues and more than 1600 question and answers, and its small size presents a realistic low-

Section: Edorta Jimenez, Biografia (*biography*)

STUDENT: **Non jaio zen Edorta Jimenez?**
Where was Edorta Jimenez born?
TEACHER: Ez dakit. Barkatu!
I don't know. Sorry!
STUDENT: **Zertara dedikatzen da Edorta?**
What does Edorta do for a living?
TEACHER: Idazketa ez ezik, itzulpena, irakasletza, kazetaritza, telebistako zein zinemako gidoilaritza izan ditu ogibide, bestek beste
Not only as a writer, but he also worked in translation, teaching, journalism, screenwriting for television and film, among others.
STUDENT: **Ze ikasketa egin ditu?**
What did he study?
TEACHER: Irakasle ikasketak egin zituen
He studied teaching
STUDENT: **Zein da bere ezizena?**
What's his nickname?
TEACHER: Omar Nabarro
Omar Nabarro
STUDENT: **Seme-alabak ditu?**
Does he have any children?
TEACHER: Bai. Irati Jimenez idazlearen aita da
Yes. Is the father of the writer Irati Jimenez
STUDENT: **Zein dira bere eleberri aipagarriak?**
What are his notable novels?
TEACHER: "Kilkerren hotsak" (2003) eta "Sukar ustelaren urtea" (2004)
"Kilkerren hotsak" (2003) and "Sukar ustelaren urtea" (2004)

Figure 1: An example dialogue in Basque (with its translation to English) where the student asks questions after reading a small introduction about the person, but without seeing the section text. The teacher answers the questions selecting a span of text of the section, adding optionally "Yes" or "No" to it, or choosing "I don't know" option.

resource scenario for CQA systems. Figure 1 displays an example of a dialogue related to a Wikipedia section about the biography of Edorta Jimenez, a Basque writer, with translations into English.

Current CQA systems rely heavily on pre-trained language models such as ELMO (Peters et al., 2018), BERT (Devlin et al., 2019) and derived models (Liu et al., 2019b; Lample and Conneau, 2019). These models are first trained on large numbers of text using a language model loss, and then fine-tuned on the train data of a CQA dataset. The most recent models include multilingual versions (Devlin et al., 2019; Lample and Conneau, 2019), where the text in different languages is represented in a common space. The multilingual versions allow to transfer trained models to languages other than that used to fine tune them.

In this paper we test the performance of variants of BERT in low-resource scenarios: native training data only, zero-shot transfer (English training data only) and low resource transfer (native and English training data). The main point of comparison with the state of the art is obtained with the publicly available multilingual BERT which covers 104 languages (Devlin et al., 2019). Given the issues with smaller languages in multilingual BERT, and in order to get competitive native results, we trained a multilingual BERT for English, Spanish and Basque and we used a monolingual version of BERT. The best results are obtained in the low resource transfer scenario. The F1 score is comparable to those reported for the English QuAC dataset.

The contributions of our work are the following: (1) We release ElkarHizketak, a low resource conversational question answering dataset in Basque, constructed in a challenging setting: unavailability of crowdworkers, and smaller Wikipedia. (2) We present the results of monolingual and multilingual BERT pre-trained language models in three settings (native training, zero-shot transfer and low resource transfer) showing that transfer from English is successful, and combined with native data produces results which are comparable to those obtained for the analogous QuAC English dataset. (3) Our experiments show that dialogue history models are not directly transferable from one language to another. The dataset is freely available with an open license[1]. To our knowledge, this is the first non-English conversational QA dataset, the first conversational dataset for Basque, and the first cross-lingual transfer results on conversational question answering.

2. Related Work

Work in conversational QA systems has led to the creation of a variety of datasets for the task (Nguyen et al., 2016; Rajpurkar et al., 2016; Iyyer et al., 2017; Trischler et al., 2017; Kočiský et al., 2018; Dunn et al., 2017). MS MARCO (Nguyen et al., 2016), NewsQA (Trischler et al., 2017) or SearchQA (Dunn et al., 2017) are some examples of reading comprehension datasets that require systems to understand a document to properly answer the queries. SequentialQA (Iyyer et al., 2017) comprises more than 6.000 question sequences where each question refers and refines previous ones, and therefore can be seen as different turns in a dialogue. More similar to our work, CoQA (Reddy et al., 2018) and QuAC (Choi et al., 2018) are two datasets that contain QA information-seeking dialogues about different topics. CoQA contains 127K questions/answer pairs from 8K conversations about passages from several domains, and QuAC contains around 14K information-seeking dialogues (100K questions in total) about people in Wikipedia. These datasets were created by crowdsourcing in a wizard-of-oz fashion. One worker (the student) was presented with an initial paragraph describing some aspect of the subject of interest and posed the initial query. A second worker (the teacher) had at his disposal a passage and had to highlight the relevant excerpt to answer the query.

In spite of being very valuable resources to build conversational QA systems, all current conversational QA datasets are in English, which makes it difficult to acknowledge progress in other languages. Research in related areas such as question answering have produced multilingual datasets such as XQA (Liu et al., 2019a), a multilingual dataset in 9 languages for open domain QA. However, no such alternative exists in the conversational QA field; as far as we know, ElkarHizketak is the first attempt to create a conversational QA dataset in a language other than English.

Contextualized word embeddings are representations that are sensitive to the context where the word appear. These models are first pre-trained on big corpora using a language modeling loss. The pre-trained model is then fine-tuned to the task at hand, using manually annotated datasets and appropriate loss functions. They have been successfully used in a variety of natural language processing tasks, including QA and dialogue systems (Devlin et al., 2019; Qu et al., 2019). ELMO (Peters et al., 2018) and Flair (Akbik et al., 2018) are language models built upon LSTM-based architectures. BERT (Devlin et al., 2019) is a model based on a transformer architecture, and pre-trained using a masked language model objective. BERT has been very successful on many NLP tasks, and several variants exists, such as RoBERTA (Liu et al., 2019b) and ALBERT (Lan et al., 2020). These models are trained for English, but some authors have built pre-trained models for other languages such as French (Martin et al., 2019). Interestingly, the knowledge learned by pre-trained models such as BERT has been shown to be transferable across domains. For instance, in Campos et al. (2019) the authors use BERT to build a conversational QA system based on FAQs. The best results are obtained using a pre-trained BERT model which is fine-tuned on QuAC, and then fine-tuned again using a much smaller FAQ dataset.

Masked language models have been extended to a multilingual setting by building a shared model that is trained with corpora in many languages. Multilingual BERT, or mBERT[2], is simultaneously trained on 104 different languages using monolingual Wikipedia data. XML (Lample and Conneau, 2019) is jointly trained on 100 languages using a masked language model objective, also including parallel corpora when available. These multilingual mod-

[1] http://ixa.si.ehu.es/node/12934

[2] https://github.com/google-research/bert/blob/master/multilingual.md

els allow to perform knowledge transfer among languages. For instance, in Artetxe et al. (2019) the authors train a multilingual BERT model and show that language knowledge transfer is helpful for cross lingual natural language inference or question answering. In this paper we apply knowledge transfer across languages and show that learning over English QuAC yields better results when tested on Basque ElkarHizketak. Moreover, we also show that the in-house built multilingual BERT, which includes only three languages, is better than using mBERT.

3. Dataset Creation

This section begins by describing the selection process of the passages to be used in the interactive task for dialogue collection described afterwards.

3.1. Passage Selection

Our passage selection process is more or less identical to the one used for QuAC dataset. We selected sections of Wikipedia articles about people, as Choi et al. (2018) indicated that less specialized knowledge is required to converse about people than other categories. In order to retrieve articles we selected the following categories in Basque Wikipedia: *Biografia* ('Biography' is the equivalent category in English Wikipedia), *Biografiak* ('People') and *Gizabanako biziak* ('Living people'). We applied this category filter and downloaded the articles using a querying tool provided by the Wikimedia foundation[3]. Once we retrieved the articles, we selected sections from them that contained between 175 and 300 words. These filters and threshold were set after some pilot studies where we check the adequacy of the people involved in the selected articles and the length of the passages in order to have enough but not to much information to hold a conversation.

3.2. Dialogue Collection

Dialogues were collected during some online sessions that we arranged with Basque speaking volunteers. We adapted the CoCoA dialogue framework (He et al., 2017) to use it as a tool for dialogue collection through a text-based chat interface in those online sessions. This interface allowed us to pair up two volunteers, who play the roles of a student and a teacher, to converse about a specific section of a Wikipedia article (such as "Biography" section of Edorta Jimenez article in the example shown in Figure 1).
Both participants can see a chatbox with the dialogue they are holding on the interface. However, the rest of the content on the interface is different for each one of them. The role of the students is to ask free text questions to the teachers, so they only need to see the title of the Wikipedia article (which is in fact the name of the person of interest), the first paragraph of the article (which is usually a brief biography of the person), and the heading of one section of the article (they should ask questions just about this section). Note that students do not see the actual content of the section, so that the actual conversation is not guided by the information represented in the content. If the answer for a question posed by the student can not be answered by looking at the passage, the question will be marked as unanswerable. The role of teachers is to answer the questions by selecting an adjacent span from the section text they are provided with. The selected text span is copied automatically into an answer box, which they can edit it to make minimal modifications and make the answer look more natural. Some restrictions are imposed to the length of both the question and the answer, which are 150 and 200 characters, respectively.
Furthermore, the teacher has to specify the following dialogue acts:

- Affirmation. It is required when the question is a Yes/No question: *yes*, *no* or *neither*.
- Answerability. It will define if the question has an answer or not: *answerable* or *no answer*. When no answer is selected, the returned string is "*Ez dakit. Barkatu!*" ("I don't know. Sorry!").

Regarding dialogue acts, we used the ones proposed in QuAC, but we removed the *continuation* act as we thought it was confusing for users, and we wanted to make the task as simple as possible to volunteers.
The student can decide to end the dialogue at any moment once the dialogue have at least 2 question-answer pairs and at least one of the answers is not "I don't know. Sorry!". If not, they will continue conversing until the dialogue has a maximum of 8 question-answer pairs, 3 unanswerable questions have been asked, or 10 minute time limit is reached. Because in such a case, the conversation will end automatically.

4. Dataset Analysis

In this section the collected data is analyzed from different points of view and it is compared to the QuAC dataset.

4.1. Overall Statistics

Some statistics of the dataset divided into training, development and testing splits are presented in Table 1, together with the overall statistics of QuAC. The splitting looks sensible as the differences among them are minor and every sections are different in the dev and test splits. The figures shown a clear difference in the size of both datasets. The average tokens per question is slightly lower in ElkarHizketak than in QuAC. But the figures revealed a significant difference in the average tokens per answer and the number of questions per dialogue. The amount of unanswerable questions is also considerably higher in ElkarHizketak. A manual inspection on such questions showed that in many cases the student did not ask about the specified section, but they asked general questions like "when did he born?" or "where did she born?".

4.2. Question Types

The most frequent initial words in the questions are *what, which, who, where, what, when* and *how many* (see Table 2). A similar pattern of questions was obtained in QuAC. As we can see in the examples, most of the questions are factoid. This implies the short length of the answers as noted in the previous section.

[3] http://petscan.wmflabs.org

	train	dev.	test	overall	QuAC
questions	1,306	161	167	1,634	98,407
dialogues	301	38	38	377	13,594
unique sections	281	38	38	357	8,854
tokens / questions	5.30	5.16	4.77	5.23	6.5
tokens / answers	7.27	6.85	7.06	7.19	14.6
questions / dialogue	4.34	4.24	4.39	4.33	7.2
extractive %	57.65	54.66	59.28	57.39	-
abstractive %	42.34	45.34	40.72	42.60	-
yes/no %	19.68	24.22	19.16	20.0	25.8
unanswerable %	32.47	28.57	28.74	31.63	20.2

Table 1: Statistics of the ElkarHizketak dataset, compared to QuAC.

bigram prefix	%	example
Zein / *what, which, who*	16.79	
da / *is*	49,10	Zein da bere ezizena? / *What's his/her nickname?*
urte / *year*	3.97	Zein urtetan joan zen Europara? / *Which year did he/she go to Europe?*
Non /*where*	7.52	
jaio / *born*	35.48	Non jaio zen Dionisio Amundarain? *Where was Dionisio Amundarain born?*
bizi / *live*	25.81	Non bizi da gaur egun? / *Where does he/she live nowadays?*
Zer / *what*	7.27	
ikasi / *to learn, to study*	15.83	Zer ikasi zuen Arturok? / *What did Arturo study?*
da / *is*	11.67	Zer da feminismo liberala? / *What is liberal feminism?*
Noiz / *when*	7.15	
jaio / *born*	25.42	Noiz jaio zen Katherine Johnson? / *When was Katherine Johnson born?*
hasi / *to begin, to start*	20.34	Noiz hasi zen Enkarni Genua literatur munduaz interesatzen? *When did Enkarni Genua become interested in the literary world?*
Ze / *what*	3.58	
ikasketak / *studies*	28.81	Ze ikasketa egin zituen? / *What did he/she study?*
urtetan / *in year*	6.78	Ze urtetan jaio zen? / *In which year was he/she born?*
Zenbat / *how much, how many*	3.58	
urte / *year*	22.03	Zenbat urtez aritu zen politikan? / *How long has he been in politics?*
seme-alaba / *children*	3.39	Zenbat seme-alaba ditu? / *How many children does he/she have?*

Table 2: The most frequent initial words and phrases of the questions of ElkarHizketak.

During a manual inspection of the dialogues of the dataset we found that some of the questions are dependent on the dialogue history, that is, it is required coreference resolution as there are some entities or events that refer to previous questions or answers in the dialogue. The dialogue example displayed in Figure 2 shows such dependence of the history. For example, in the second question the student is asking about the movie that won the prize that was mentioned in the previous answer (i.e., the name of the prize is omitted in the current question). In the following question the name of the movie is omitted, so a back reference to the previous is needed in order to know which movie are they asking about. Later in the dialogue there is again a coreference as they asked about any other works than the film mentioned previously.

5. Experimental Setting

In this section we present the task definition and the baseline models.

5.1. Task Definition

Given a question and a passage as an input, conventional QA systems are designed to find a relevant excerpt in the passage which answers the question. These systems have evolved in a trickier CQA systems which have to handle with a sequence of questions which might be dependent among them. In other words, in order to fully understand the current question it might be needed to take into account the dialogue history as it could have references to previous questions or answers. Thus, CQA systems take also as an input the dialogue history which consists of previous question/answer pairs. Moreover, the system presented in this paper has to predict yes/no answer dialogue acts as an output, which are needed for affirmation questions.

Thus, we define the task with the following inputs: current question q_k, the answer passage p and the dialogue history $\{q_1, a_1, ...q_{k-1}, a_{k-1}\}$ which consists of questions and respective ground truth answers. And the outputs will be the answer span a_k with the i starting index and j ending index as boundaries in the passage p, and dialogue act list v, which will contain $\{$*yes,no,-*$\}$ values for predicting affirma-

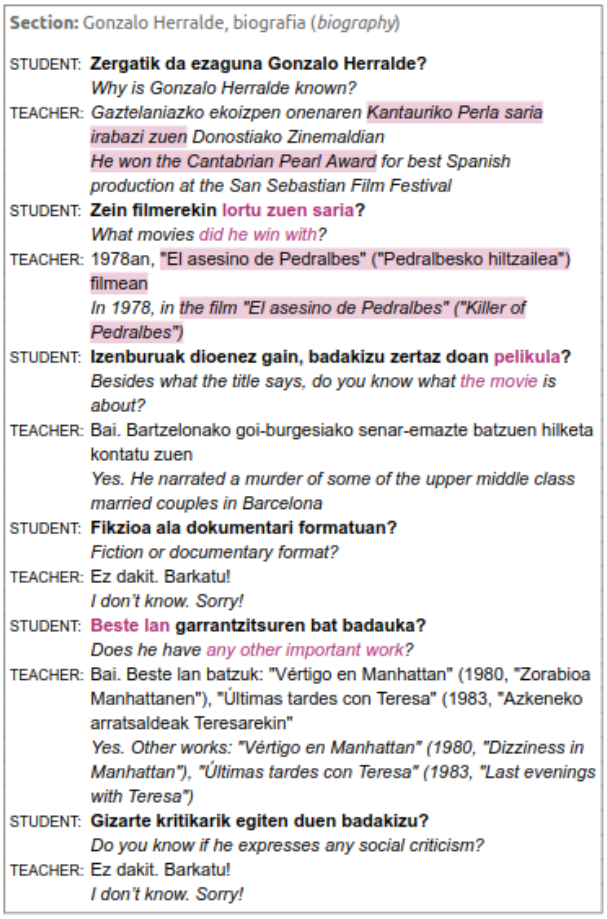

Figure 2: An example of a dialogue where there are many references in the questions to previous answers in the dialogue.

tion.

5.2. Baseline Models

In this section we present the different baseline models we have developed for the ElkarHizketak dataset. The first one is a simple majority class baseline. The following three baselines are based on language models that do not take into account any dialogue history, while the last three models do.

`Majority:` The majority answer baseline always returns "*ez dakit*" ("I don't know").

`mBERT:` A multilingual language model pre-trained simultaneously on the Wikipedia articles of 104 different languages released by Devlin et al. (2019). In all of our experiments we use the $\text{mBERT}_{\text{BASE}}$ configuration for fair comparison between the different models.

`BERTeus:` We have used the pre-trained BERT model for the Basque Language (Agerri et al., 2020) due to the low representation this language has in the official multilingual BERT model. This Basque BERT model has been trained on a corpus comprising the Basque Wikipedia and news articles from Basque media.

`mBERT_ours:` We have pre-trained a multilingual BERT model with the intention of performing transfer experiments from high resources languages as English and Spanish to Basque. This transfer experiments could be already performed with the official `mBERT` model, but as it covers that many languages, Basque is not very well represented. In order to create this new multilingual model that contains just English, Spanish and Basque, we have followed the same configuration as in the `BERTeus` model. We re-use the same corpus of the monolingual Basque model and add the English and Spanish Wikipedia with 2.5M and 650M tokens respectively. Due to the imbalance of the input corpora sizes, we have used the same oversampling and sub-word vocabulary creation strategies proposed in Lample and Conneau (2019). At the end, we have a multilingual sub-word vocabulary of 112K tokens.

The previous models do not handle any dialogue context. In contrary, for the following baseline models, we chose a History Answer Embedding (HAE) approach for BERT-based models introduced by Qu et al. (2019) for **dialogue history modeling**. The system includes dialogue history $\{q_1, a_1, ...q_{k-1}, a_{k-1}\}$ to BERT by adding a history answer embedding that marks if a token is part of history or not to other embeddings. The three systems are the following:

`BERTeus + HAE:` HAE built on top of the `BERTeus` model.

`mBERT + HAE:` HAE built on the `mBERT` model.

`mBERT_ours + HAE:` HAE built on the `mBERT_ours` model.

6. Evaluation

In this section we present the evaluation metrics, the experimental setup and the results.

6.1. Evaluation Metrics

F1 is the main evaluation metric and is computed by the overlap at word level of the prediction and the reference answer. Note that as contrary to QuAC, the test set of ElkarHizketak does not contain multiple answers for each question, so only one F1 score is provided (F1 score computed after filtering out answers with a low agreement was also provided in QuAC).

6.2. Experimental setup

All the experiments were carried out using the extractive information of the train/dev/test splits of ElkarHizketak. The baselines that use the monolingual `BERTeus` model are trained and evaluated using only the ElkarHizketak dataset (**native training**). Regarding the cross-lingual models, apart from the just mentioned approach, another two different cross-lingual transfer learning approaches are followed:

- **zero-shot** cross-lingual transfer: we use the train data of QuAC for training the model, and evaluate it on ElkarHizketak.
- **low resource** cross-lingual transfer: once we have the previous model, we fine tune it using the small train split of ElkarHizketak and test it on ElkarHizketak test split. For completeness, both development and test figures are shown.

	native training		zero-shot transf.		low resource transf.	
Model	dev.	test	dev.	test	dev.	test
Majority	28.6	28.7	28.6	28.7	28.6	28.7
without dialogue history						
BERTeus	32.4	35.0	-	-	-	-
mBERT	28.8	28.8	31.2	31.5	37.0	37.4
mBERT_ours	31.8	35.7	38.5	38.9	42.7	**41.2**
with dialogue history						
BERTeus + HAE	39.4	**40.1**	-	-	-	-
mBERT + HAE	30.7	31.4	28.3	33.3	33.0	28.7
mBERT_ours + HAE	41.2	37.4	37.0	**40.7**	43.0	40.0

Table 3: F1 scores of the baseline models in three different settings: native train and testing on ElkarHizketak (columns 2 and 3), zero-shot transfer learning where QuAC train split is used for training and it is evaluated on ElkarHizketak (columns 4 and 5), and low resource transfer learning where the previous model is fine tuned using the small train set of ElkarHizketak (columns 6 and 7). Best results on test for each scenario in bold.

6.3. Results

Table 3 shows the F1 scores obtained by all models (including models that do not handle dialogue history and the ones that do handle it) in three different settings (native training, zero-shot transfer learning and low resource transfer learning).

The majority class baseline underperforms all models in all settings, except mBERT (with and without HAE) as in some of the settings both achieve similar results.

Regarding the three models that do not handle dialogue history, the results demonstrate the validity of cross-lingual transfer learning from English, as we improve the results of native training in both transfer settings, and low-resource transfer learning approach yields increasingly good results on data. The best results are obtained using our in-house multilingual BERT, which beats the official mBERT in all three scenarios and yields results comparable to the monolingual BERTeus on the native scenario (slightly worse on development data, slightly better on test).

The results when modeling dialogue history using HAE show an unexpected pattern. In the native scenario, all models get a significant improvement when adding dialogue history, and again, BERTeus and our multilingual BERT perform comparably (slightly better on development data, slightly worse on test), and better than the official multilingual BERT. The transfer learning scenario, though, does not show improvements for the use of dialogue history. In the zero-shot transfer scenario, none of the multilingual BERT models shows improvement in both development and test (only in test). In the low resource transfer scenario the official multilingual BERT degrades when using HAE, while the results of our multilingual BERT are comparable (slightly better for development, slightly worse for test). All in all, our results show that, contrary to the BERT models fine-tuned for QA, the HAE subcomponent cannot be transferred to another language straightaway. We leave research on solutions for transferable dialogue history models for the future.

Note that the best results on the Basque dataset are roughly comparable to those reported for the English QuAC dataset[4], which shows that cross-lingual transfer is successful also in low-resource regimes.

7. Conclusions and future work

To sum up, we have presented ElkarHizketak, a low resource conversational question answering dataset in Basque, constructed in a challenging setting: unavailability of crowdworkers, and smaller Wikipedia. It is the first non-English CQA dataset and it is publicly available. We have studied the performance of baseline CQA systems in three settings: native training, zero-shot transfer from English and low resource transfer (combination of transferring the English model and combining it with the native training data). The best results are obtained in the last scenario, with results comparable to those reported in an analogous dataset for English, QuAC, showing that it is possible to obtain good results with low amounts of native data thanks to cross-lingual transfer learning. We also show that dialogue history models are not directly transferable from one language to another. For the future, we plan to research on transferability of dialogue models across languages.

8. Acknowledgements

This research was partially supported by ERA-Net CHIST-ERA LIHLITH Project funded by the Agencia Estatal de Investigación (AEI, Spain) project PCIN-2017-118, the 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 IXA excellence research group), 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. We also acknowledge the support of Google Cloud.

[4]The results on QuAC for BERT reach an overall F1 of 54.2 when using multiple reference answers (Qu et al., 2019) and of 39.2 when using a single reference answer. The results on ElkarHizketak use a single reference answer.

9. Bibliographical References

Agerri, R., San Vicente, I., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., and Agirre, E. (2020). Give your text representation models some love: the case for basque. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Artetxe, M., Ruder, S., and Yogatama, D. (2019). On the Cross-lingual Transferability of Monolingual Representations. *arXiv preprint arXiv:1910.11856*.

Campos, J. A., Otegi, A., Soroa, A., Deriu, J., Cieliebak, M., and Agirre, E. (2019). Conversational QA for FAQs. *To appear in 3rd Conversational AI Worksop at the NeurIPS Conference*.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). QuAC: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

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

He, H., Balakrishnan, A., Eric, M., and Liang, P. (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.

Iyyer, M., tau Yih, W., and Chang, M.-W. (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.

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. (2018). The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Liu, J., Lin, Y., Liu, Z., and Sun, M. (2019a). XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368.

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

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and Sagot, B. (2019). CamemBERT: a Tasty French Language Model. *arXiv preprint arXiv:1911.03894*.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv preprint arXiv:1611.09268*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *NAACL*.

Qu, C., Yang, L., Qiu, M., Croft, W. B., Zhang, Y., and Iyyer, M. (2019). BERT with History Answer Embedding for Conversational Question Answering. In *SIGIR '19*.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Reddy, S., Chen, D., and Manning, C. D. (2018). CoQA: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. (2017). NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.