information

MDPI

Article

Exploring the Utility of Dutch Question Answering Datasets for Human Resource Contact Centres

Chaïm van Toledo [1,*], Marijn Schraagen [1], Friso van Dijk [1], Matthieu Brinkhuis [1] and Marco Spruit [2]

[1] Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands
[2] Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
* Correspondence: c.j.vantoledo@uu.nl

Abstract: We explore the use case of question answering (QA) by a contact centre for 130,000 Dutch government employees in the domain of questions about human resources (HR). HR questions can be answered using personnel files or general documentation, with the latter being the focus of the current research. We created a Dutch HR QA dataset with over 300 questions in the format of the Squad 2.0 dataset, which distinguishes between answerable and unanswerable questions. We applied various BERT-based models, either directly or after finetuning on the new dataset. The F1-scores reached 0.47 for unanswerable questions and 1.0 for answerable questions depending on the topic; however, large variations in scores were observed. We conclude more data are needed to further improve the performance of this task.

Keywords: question and answering; dataset; Squad 2.0; question and answering dataset creation; question and answering Dutch; human resource dataset

check for updates

Citation: van Toledo, C.; Schraagen, M.; van Dijk, F.; Brinkhuis, M.; Spruit, M. Exploring the Utility of Dutch Question Answering Datasets for Human Resource Contact Centres. *Information* **2022**, *13*, 513. https://doi.org/10.3390/info13110513

Academic Editor: Kostas Stefanidis

Received: 9 September 2022
Accepted: 23 October 2022
Published: 28 October 2022

1. Introduction

In a contact centre (CC), human resources (HR) employees need to answer often complex questions from organisational users that necessitate access to personal files and a thorough understanding of HR documentation. However, much time is spent answering simple questions. Substantial time would be saved if we could handle the simple questions automatically. Therefore, we need to determine whether a computer can handle a particular question or not.

For a long time, researchers have tried to let a computer answer questions. In recent years, enormous datasets have been publicly available to let computers learn how to answer questions. In Squad 2.0 [1], the dataset's structure is modelled to learn how to answer questions and determine whether a particular question cannot be answered.

This research will examine whether this technique can be employed in a CC at the Dutch government. Our case study organisation is P-Direkt; it has a CC of more than 425 full-time employees. Questions cover policies, such tax benefits for bicycle commuting and how the HR portal works.

Every year around 240,000 questions are registered from 130,000 users, creating a very high workload. With the end goal of reducing this workload, we investigated the following research questions: Can Dutch question answering (QA) datasets help CCs answer questions automatically? How do modelled HR questions perform on current Dutch QA models? How well does a self-trained QA model perform on answering CC questions? What kind of dataset varieties support modelling our Dutch HR QA dataset?

One important constraint is that the Dutch QA community is much smaller than the English community. Currently, there are no suitable publicly available Dutch QA datasets. We investigated whether we could employ machine-translated QA datasets in Dutch.

Therefore, our contributions include how to make a small QA dataset from e-mail. Another contribution was to explore how a small organisation-specific QA dataset performed, as opposed to a sizeable general-purpose QA dataset developed using crowdsourcing platforms such as Amazon Mechanical Turk.

The paper is structured as follows: in Section 2, entity linking, QA datasets, and translated datasets are elaborated. In Section 3, we describe the creation of the P-Direkt QA dataset. Section 4 explains how the analysis was performed. Section 5 reports the details of the created dataset and the test statistics, followed by discussion and suggested future work in Sections 5.3 and 6. We answer our research questions in Section 7.

2. Background

The first part of the background discusses how entity-linking methods work, which are needed to create a QA dataset. The second part introduces QA and their datasets.

2.1. Entity Linking

Rao et al. [2] described three entity linking pipelines: (1) entity recognition (NER), (2) coreference resolution, and (3) relation extraction. The task of NER is to identify bounded entities, such as persons and organisations. Coreference resolution involves linking entities that are not written identically. For example, a country named Holland, the Netherlands, Pays-Bays, and Nederland. The task of relation extraction is to identify relationships to documents, such as identifying the same person in different news articles.

2.2. QA and QA Datasets

QA is a hot topic within the natural language processing (NLP) domain. In recent years, many crowdsourced public datasets have become available. These datasets have been used for machine learning approaches to QA, with results that often match or outperform a human performance [3].

Numerous QA datasets have also been publicly presented, including NewsQA [4], Squad [5], QuAC [6], Coqa [7], and more. There are also closed-domain datasets, such as RecipesQA for cooking recipes [8], the IBM TechQA dataset [9], and JEC-QA [10]. Another task is to translate these datasets into different target languages, as most of these datasets were crowdsourced in English.

Due to the limitations of the language boundaries, many attempts have been made to expand existing QA datasets into other languages. For example, the Squad dataset has been translated into Spanish [11], French [12], Persian [13], Arabic [14], and other languages.

Rogers et al. [15] distinguished two tasks of datasets, an information-seeking task and a probing task. With information-seeking, the task comes from users needing information. With probing questions, the task is to create questions from a source text to create the dataset.

3. Dataset

In this study,we needed to create a dataset to examine whether the data structure of the Squad 2.0 dataset was suitable for questions in the HR domain. The starting point was a dataset from approximately 170,000 P-Direkt e-mails and 295 HR documentation pages from the Dutch government intranet.

The e-mails were sent between 2018 and 2020. The personal identifiers in the e-mails were morphed into pseudonymised placeholders [16]. The intranet documentation was scraped with Scrapy [17] in May 2021 and cleaned with Beautiful Soup [18].

The questions in the e-mails showed us that every question was answerable with either documentation, personnel files, or both. We focused only on the questions with an answer from documentation. Our first exploration with a subset of 202 e-mails showed that 18% of the questions could be answered from written documentation.

We created an answerable question identification dataset that distinguished between questions answerable by documentation and those that were not by randomly selecting sentences with a question mark, each with at least six tokens or more. From a random selection of 5409 questions, we categorised 820 questions as valid, i.e., answerable from documentation. These questions were not linked to answers and were only used to identify questions for the QA dataset creation.

Figure 1 gives a brief overview of the construction of the dataset. The first step is selecting the correspondence. With this answerable question identification dataset, we created a model with the Roberta For Sequence Classification algorithm with the Dutch language model RobBERT [19]. We could therefore identify well-written questions with an information need based on documentation instead of their personal files. These correspondences were chosen to analyse further.

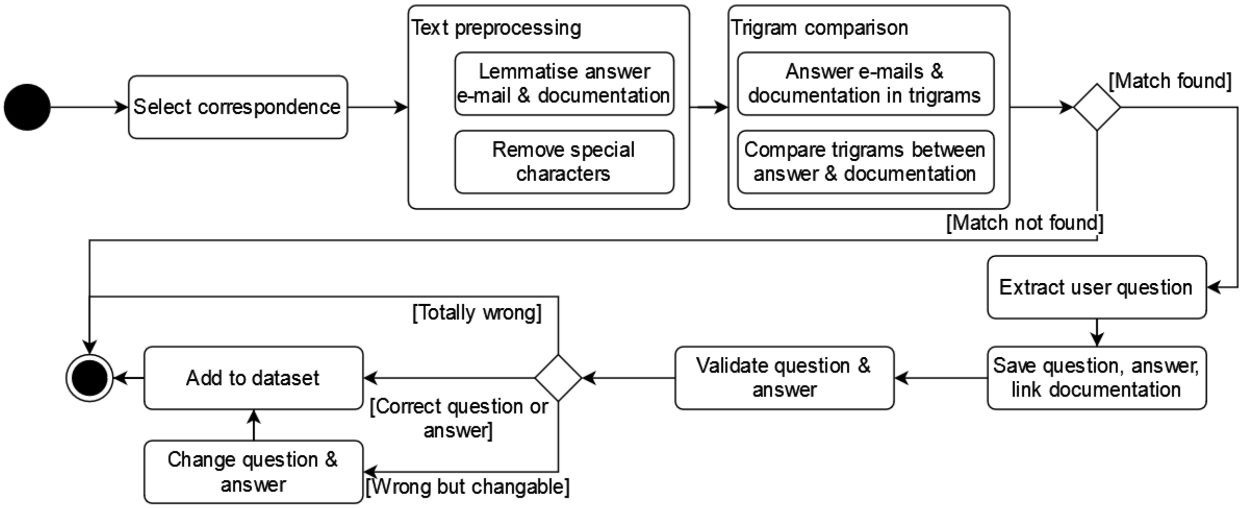
Select correspondence
Text preprocessing
Lemmatise answer e-mail & documentation
Remove special characters
Trigram comparison
Answer e-mails & documentation in trigrams
Compare trigrams between answer & documentation
[Match found]
[Match not found]
Extract user question
Save question, answer, link documentation
Validate question & answer
[Totally wrong]
[Correct question or answer]
[Wrong but changable]
Add to dataset
Change question & answer

Figure 1. Dataset creation steps.

In the second step, the preprocessing was performed for the texts. For both answer e-mails and documentation, we removed stopwords and non-alphabetical tokens and lemmatised the remaining tokens using the spaCy toolkit [20].

The third step was to find a link between the answers and the documentation. A link must have an exact match between the answer and documentation [21]. This was performed by forming the texts in trigrams and finding comparisons between the answer e-mails and the documentation.

The fourth step extracted the question from the customer's e-mail. A straightforward method found the question. The method tokenised the sentences and selected the first sentence with a question mark as the leading question [22]. In the fifth step, the question, answer and documentation were combined in a JSON file, as an invalidated temporary dataset.

After data mining the e-mail dataset, the validation began, at step six. The tokenisation process did not always give clear sentences. Sometimes, the questions had noise from previous sentences and incomplete tokenisation (substep change question and answer, Figure 1). We cleaned the noisy questions to their essence. The same applied to the documentation. Sometimes, the documentation link had headings or was too extensive and had unnecessary information. We also cleaned those answers to their essence. We noticed that not every paragraph could be connected to an answer. Therefore, we added 30 additional questions with answers from the documentation.

To mimic the Squad 2.0 dataset, our data was stored in JSON. A key difference with Squad 2.0 was our use of a helpdesk e-mail archive as a source, as opposed to Squad 2.0, which was a crowdsourced approach to create artificial fact-based questions.

This dataset came in three different versions: full, medium, and small. Figure 2 gives a global overview of the differences between the three versions. The full dataset contained all the found questions and answers divided over six pages of HR-related texts, for six documents. The medium dataset contained questions with unique answers but only answerable questions. The unanswerable questions were the same in number as the full dataset. The small dataset contained questions with unique answers and unanswerable questions linking to unique plausible answers.

Figure 2 gives a brief overview of the construction of the dataset. The first step was to find a link between the answers and the documentation. A link must have an exact match between the answer and documentation [21]. The next step extracted the question from the customer's e-mail.

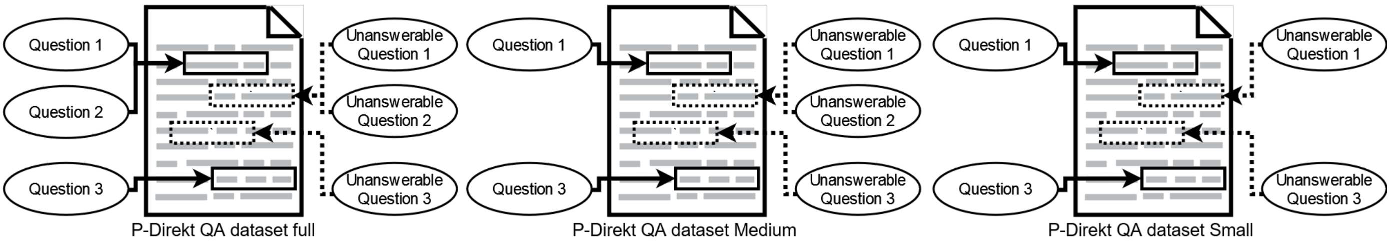

Figure 2. Visualisation of the P-Direkt question answering (QA) datasets: full, medium, and small.

Table 1 represents the transformation of the dataset. Before comparing the answer and the documentation, a few transitions occurred in the e-mail dataset and documentation. The datasets are available online (https://github.com/7083170/Exploring-the-utility-of-Dutch-question-answering-datasets-for-human-resource-contact-centres, accessed on 18 October 2022).

Table 1. Example of question and answer extraction with the result. Translated automatically from Dutch to English with Google.

	Question E-Mail	Answer E-Mail CC	Documentation
Original data	Good afternoon, I have a question about the PAS scheme because I cannot find the answer anywhere. If you are sick, do the PAS hours expire? I have scheduled the PAS hours every week for the whole year. [SIGNATURE]	Dear [PERSON], You contacted P-Direkt on [DATE]. You would like information about absenteeism if you use the PAS scheme. Sick and Pass. The time you do not have to work because of the PAS scheme is not working time, nor is it a leave. Are you sick on a so-called PAS day? Then you cannot take these hours at another time. You were ill at a time when you did not have to work. For more information, see the Government Portal A to Z list [...]	Consequences of PAS regulation [...] Sickness The time you do not have to work because of the PAS scheme is not working time, nor is it a leave. Are you sick on a so-called PAS day? Then you cannot take these hours at another time. You were ill at a time when you did not have to work. [...]
Extracted data	If you are sick, do the PAS hours expire?	The time you do not have to work because of the PAS scheme is not working time, nor is it a leave. Are you sick on a so-called PAS day? Then you cannot take these hours at another time.	[idem]

4. Method

The questions in the dataset could be answered using documents on six different topics (see Table 2). For analysis, six-fold cross-validation was used to train the models on each combination of six documents, and we tested the model on the remaining document in each fold, modelled in Figure 3. We reported the F1 score as the evaluation metric.

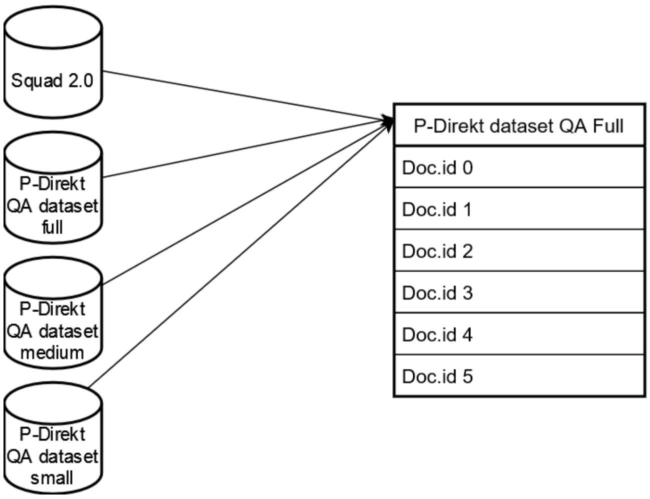

Figure 3. Sixfold cross validation with four different training datasets.

We used the BERT language model architecture. Since the rise of BERT [23], language models have played a significant role in different kinds of NLP tasks, such as sentiment analysis, QA, NER, and others. The BERT models outperform traditional approaches, especially the RoBERT variant [24]. For QA, the RoBERTa model has performed better than the other BERT variants [25]. The Dutch language has a few variants of the BERT language models, namely, BERTje [26], RobBERT [19], BERT-NL [27], and multilingual BERT [23]. In the experiment, we used the RobBERT variant.

Table 2. Extracted questions with answers per document. Translated automatically from Dutch to English with Google.

Doc.id	**Dutch Original Title**	**English Meaning**	**Number of Questions**		
			Full	**Medium**	**Small**
0	IKB uren en spaarverlof	Flexible vacancy hours and savings hours leave	100	36	31
1	Fiets	Bicycle	66	51	49
2	PASregeling neveninkomsten	Partial employment rate for seniors	38	24	22
3	Minuren	Negative vacancy hours	50	25	20
4	Betaalbewijzen	Payment receipts	30	22	21
5	Compensatieuren	Compensation hours	38	34	14
		Total:	322	192	157

The three parts of the analysis used different models every time. Training and evaluation were conducted with a Huggingface script [28] (https://github.com/huggingface/transformers/blob/ec07eccc7d61a77ca3c0463f67bcde18b9943fea/examples/legacy/question-answering/run_squad.py, accessed on 1 July 2022). Testing was calculated with the F1-score.

We first implemented an off-the-shelf model trained and finetuned on Squad 2.0 data that was automatically translated to Dutch [29]. This model was trained on a BERT Base Multilingual cased model on two epochs, with a max sequence length of 384 and a doc stride of 128. The exact match score was 67.38, and the F1 score was 71.36. The tasks of giving answers and not giving answers were not in balance. HasAns (ExactMatch/F1) had 47.42 and 57.76, while the NoAns F1 score was 79.88.

The second experiment used the RobBERT model [19], finetuned on the full P-Direkt dataset using the same hyperparameters as in the first experiment. Some of the questions linked to the same answer.

The third experiment used the same setup with the P-Direkt QA medium dataset to investigate the aspect of answerable versus unanswerable questions. However, the answerable questions were still the same in number as in the full dataset. Hence, this part used the P-Direkt QA dataset medium.

The fourth experiment uses the same setup as the second experiment for the P-Direkt QA small dataset. This experiment investigated the mapping of multiple questions per answer (full dataset) versus a single question per answer (small dataset).

5. Results

We discuss the dataset in comparison with the Squad 2.0 dataset (Section 5.1) and the QA performance (Section 5.2).

Table 3. Examples of questions and answers per topic. Translated from Dutch to English with Google.

Doc.id	Question and Answer	Answerable
0	Is it true that my IKB hours for 2020 will expire if I don't take them? You can save your IKB hours in your IKB savings leave. IKB hours you have not taken as leave and have not had it paid out will be added to your IKB savings leave at the end of December. Your IKB savings leave cannot expire.	Answerable
1	Do I have to pay in advance for the bicycle and declare it afterwards? Please note that you must provide supporting documents with your application.	Not answerable
2	Can I go back from the PAS scheme to my old hours and salary? Even if you stop participating in the PAS scheme now, you can start participating again later.	Answerable
3	Can you tell me why I have negative compensation hours (-4.00)? Does a holiday fall on a day you should work according to your schedule? Then it may happen that you ultimately do not reach the number of hours you have to work annually. This creates a negative balance of compensatory leave.	Answerable
4	Can I apply for IKB per month for a tax-friendly purpose for two years? You can only submit an IKB application in the year in which you incurred the costs.	Answerable
5	Are 'compensation hours' the same as 'extra-statutory holiday hours'? Compensation hours are hours that you save by structurally working more than the number of hours stated in your employment contract. You can take these 'overworked' hours as free time at a later time	Answerable

5.1. Data Comparison

Table 3 shows the examples of the question and answer per topic. Doc.id refers to the doc.id in Table 2. All but one were answerable in this example table. Only the question at doc.id 1 was not answerable with the text.

Table 4 and Figure 4 compare the P-Direkt QA dataset and Squad 2.0. The ratio between answerable and unanswerable questions was similar. However, the word count in the context documents and answers differed considerably. Helpdesk employees elaborate the answers with whole sentences, while the Squad dataset contained short phrases expressing simple facts. In contrast, while the answers in Squad 2.0 were factual, the context documents were over twice as long on average. Figure 3 shows that the distribution of the dataset had approximately three words per answer.

In comparison, the P-Direkt dataset contained 24 words per answer. Nevertheless, the word count of the paragraphs, or formatted as the context in the dataset, differed. There, the paragraphs tended to be shorter in the P-Direkt dataset.

Table 4. Descriptive differences between the Squad 2.0 and P-Direkt question answering (QA) dataset.

	Squad 2.0	**P-Direkt QA Dataset**
Number of questions	150,000	322
Answerable questions	100,000	214
Not answerable questions	50,000	108
Number of paragraphs	14,873	63
Average words per context	134.33	65.11
Average word count answer	3.37	24.36

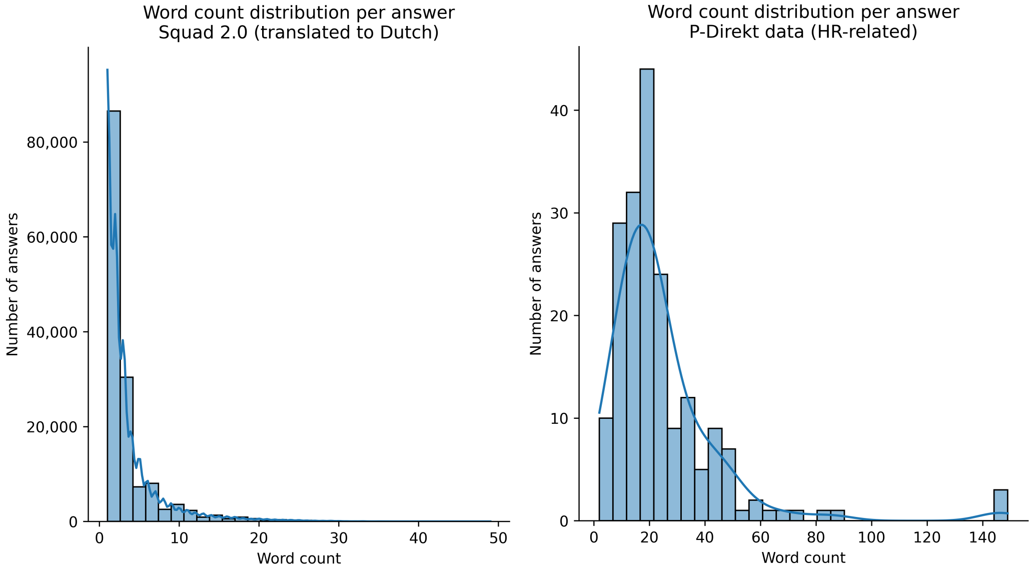

Figure 4. Distribution differences between the Dutch Squad 2.0 and P-Direkt QA dataset.

5.2. QA Performance

The test statistics are shown in Tables 5–8. The output of the not answerable questions was the same at NoAns exact as NoAns F1. Therefore, the NoAns exact match was left out of the result tables.

We start with Table 5, the test with the Squad 2.0 model. While this model accurately recognized unanswerable questions across all topics, the performance on the answerable questions was deficient at <10% F1-score and near-zero exact match. Doc.id 2, a page about negative vacancy hours, was 100 per cent. The usefulness of this model was worthless because the prediction of the answers was below 10 per cent.

Table 5. Test results from the Dutch translated Squad 2.0 model. Evaluation on the P-Direkt QA full dataset.

Doc.id	**Exact**	**F1**	**Total**	**HasAns Exact**	**HasAns f1**	**HasAns Total**	**NoAns f1**	**NoAns Total**
0	30.00	34.27	100	1.47	7.75	68	90.63	32
1	27.27	31.98	66	2.08	8.55	48	94.44	18
2	26.32	30.79	38	0.00	6.08	28	100.00	10
3	26.00	26.00	50	0.00	0.00	36	92.86	14
4	26.67	28.61	30	0.00	2.77	21	88.89	9
5	65.79	67.42	38	0.00	5.17	12	96.15	26
Mean	33.67	36.51		0.59	5.05		93.83	

Table 6. Training and evaluation on the P-Direkt QA full dataset.

Doc.id	Exact	F1	Total	HasAns Exact	HasAns f1	HasAns Total	NoAns f1	NoAns Total
0	10.00	16.89	100	8.82	18.95	68	12.50	32
1	21.21	25.51	66	0.00	5.91	48	77.78	18
2	28.95	30.86	38	3.57	6.17	28	100.00	10
3	22.00	34.67	50	2.78	20.37	36	71.43	14
4	23.33	39.79	30	0.00	23.51	21	77.78	9
5	65.79	65.79	38	0.00	0.00	12	96.15	26
Mean	28.55	35.58		2.53	12.48		72.61	

Table 7. Training on the P-Direkt QA medium dataset and evaluation on the P-Direkt QA full dataset.

Doc.id	Exact	F1	Total	HasAns Exact	HasAns f1	HasAns Total	NoAns f1	NoAns Total
0	8.00	14.04	100	7.35	16.24	68	9.38	32
1	27.27	29.63	66	2.08	5.33	48	94.44	18
2	26.32	26.32	38	0.00	0.00	28	100.00	10
3	18.00	26.87	50	0.00	12.31	36	64.29	14
4	16.67	32.39	30	9.52	31.99	21	33.33	9
5	0.00	5.37	38	0.00	17.01	12	0.00	26
Mean	16.04	22.44		3.16	13.81		50.24	

Table 8. Training on the P-Direkt QA small dataset and evaluation on the P-Direkt QA full dataset.

Doc.id	Exact	F1	Total	HasAns Exact	HasAns f1	HasAns Total	NoAns f1	NoAns Total
0	6.00	15.49	100	8.82	22.78	68	0.00	32
1	4.55	22.45	66	4.17	28.78	48	5.56	18
2	13.16	36.66	38	14.29	46.18	28	10.00	10
3	2.00	27.82	50	2.78	38.64	36	0.00	14
4	16.67	36.63	30	19.05	47.57	21	11.11	9
5	0.00	4.83	38	0.00	15.30	12	0.00	26
Mean	7.06	23.98		8.18	33.21		4.44	

Fine-tuning on the P-Direkt full dataset showed an improvement in the answerable questions, indicating that using in-domain training data was beneficial (see Table 6). However, this improvement came at the expense of the performance on the non-answerable questions. Furthermore, the improved F1-score on the answerable questions was still low at 12.5%, leading to overall lower performance. Again, doc.id 2 showed a 100 per cent score on the not answerable questions. However, the overall mean dropped by more than 20 per cent.

Training using a single question per answer (medium dataset, Table 7) showed a drop in performance for the unanswerable questions with only a slight increase in the answerable questions. Mapping unanswerable questions to plausible answers (small dataset, Table 8) dramatically improved the performance for the answerable questions. However, the unanswerable question performance showed a significant drop caused by the lack of unanswerable questions during the training.

5.3. Discussion

As mentioned before, there are no publicly available Dutch QA datasets. Prompted by the lack of suitable publicly available Dutch QA datasets, we collected a small dataset in the HR domain that allowed us to perform experiments using representative real-world data. This provided an advantage over commonly used data such as Squad 2.0, based on artificial question formulation for short fact-based answers. We note that we supplemented our dataset with around 10% of manually added questions to increase the coverage of the context documents, but these added questions were also realistic. The dataset provided an overview of how people formulated the question for desired information instead of having an instruction to formulate questions at content. The 10% of manually designed questions, were due to not finding questions that referred to certain answers in the texts. We believe that these manually designed questions are written in the same mindset as the other questions. and therefore the meaning of these questions are retained.

Another hurdle would be to add more diverse questions. In developing the current dataset, we excluded questions consisting of a statement followed by a small question: "Is this possible?" For the next version of the dataset, we will try to detect these kinds of questions and enrich these questions with previous sentences. In this iteration, we did not use these kinds of statement questions.

In the tests, the outcome varied every time. The experiments showed substantial variation in outcomes across runs. For production, the results need to stabilize as well as improve. Therefore, future research will add more content, questions, and answers to the dataset.

The Dutch-translated Squad 2.0 was not built for questions with more content. It did an excellent job on the test of unanswerable questions. However, for the answering task, it did a poor job. Still, the data structure in JSON format is worth examining further to answer employees' questions about HR. What if the questions, answers, and content double, triple, or increase tenfold? The processes in the CC should be amended by modelling the dataset to a newer and more extensive one by the helpdesk employees. On the other hand, it is also worth examining the possibilities for smaller datasets.

6. Future Research Directions

There are various possible avenues to expand on the current experiments. Our current approach isolated the questions from original e-mails to focus on the QA task. However, taking the full original context into account has shown to be beneficial in a conversational context [30,31], which may also apply to our use case of long single-turn questions.

Once we incorporate domain-specific background knowledge into the models, it becomes possible to leverage the extensive collection of e-mails for either pretraining or passage retrieval for answer formulation instead of using only the small set of official documents as a source of answers.

The performance differences for the non-answerable class prompt a more detailed investigation of this class to determine whether further subdividing this question can benefit performance. Given the length and complexity of the current answers and source documents, a further knowledge modelling or machine learning-based effort to structure the amount of information would considerably simplify the overall Q&A task in this domain.

From a business implementation perspective, it is worth pursuing different scenarios of using the models as a decision or process support system. Additionally, the business process can, in turn, support the modelling effort by incorporating data collection and labelling as part of the daily helpdesk activities with a minimum of extra effort from helpdesk personnel.

7. Conclusions

We investigated the question of whether QA datasets could help HR CCs to answer questions. We conclude that they show potential. This paper presented the P-Direkt QA dataset based on activities in their CC. The questions modelled in the dataset were real questions from users with real information needs. We argued that the way of answering questions was not the same as state-of-the-art QA datasets. Therefore, the reading comprehension was more extensive than the Squad 2.0 dataset. Furthermore, the P-Direkt dataset needs to be enlarge for better performance.

Author Contributions: Conceptualization, C.v.T.; methodology, C.v.T.; software, C.v.T.; validation, C.v.T.; formal analysis, C.v.T.; investigation, C.v.T. and F.v.D.; resources, C.v.T. and F.v.D.; data curation, C.v.T. and F.v.D.; writing—original draft preparation, C.v.T.; writing—review and editing, C.v.T., M.S. (Marijn Schraagen), F.v.D., M.B. and M.S. (Marco Spruit); visualization, C.v.T.; supervision, M.S. (Marijn Schraagen), M.B. and M.S. (Marco Spruit); project administration, C.v.T.; funding acquisition, M.B. and M.S. (Marco Spruit). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by P-Direkt, Ministry of the Interior and Kingdom Relations, The Netherlands.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The question answering dataset is available at https://data.overheid.nl/dataset/vraag-en-antwoord-dataset-rijksportaal-personeel, accessed on 22 September 2022, the different kinds of datasets are also available at https://github.com/7083170/Exploring-the-utility-of-Dutch-question-answering-datasets-for-human-resource-contact-centres, accessed on 18 October 2022.

Acknowledgments: We would like to thank the employees of P-Direkt who made the data available for doing research.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
CC	Contact Centre
HR	Human Resources
QA	Question answering
NER	Named entity recognition

References

1. Rajpurkar, P.; Jia, R.; Liang, P. Know What You Don't Know: Unanswerable Questions for SQuAD. *arXiv* **2018**, arXiv:1806.03822.
2. Rao, D.; McNamee, P.; Dredze, M. Entity Linking: Finding Extracted Entities in a Knowledge Base. In *Multi-Source, Multilingual Information Extraction and Summarization*; Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 93–115. [CrossRef]
3. Zhang, Z.; Yang, J.; Zhao, H. Retrospective reader for machine reading comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 14506–14514.
4. Trischler, A.; Wang, T.; Yuan, X.; Harris, J.; Sordoni, A.; Bachman, P.; Suleman, K. NewsQA: A Machine Comprehension Dataset. In Proceedings of the 2nd Workshop on Representation Learning for NLP, Vancouver, BC, Canada, 3 August 2017; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 191–200. [CrossRef]
5. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv* **2016**, arXiv:1606.05250.
6. Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; Yih, W.t.; Choi, Y.; Liang, P.; Zettlemoyer, L. QuAC: Question Answering in Context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 2174–2184. [CrossRef]
7. Reddy, S.; Chen, D.; Manning, C.D. CoQA: A Conversational Question Answering Challenge. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 249–266. [CrossRef]
8. Yagcioglu, S.; Erdem, A.; Erdem, E.; Ikizler-Cinbis, N. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 1358–1368. [CrossRef]
9. Castelli, V.; Chakravarti, R.; Dana, S.; Ferritto, A.; Florian, R.; Franz, M.; Garg, D.; Khandelwal, D.; McCarley, J.S.; McCawley, M.; et al. The TechQA Dataset. In Proceedings of the Association for Computational Linguistics (ACL), Seattle, WA, USA, 5–10 July 2020; pp. 1269–1278.
10. Zhong, H.; Xiao, C.; Tu, C.; Zhang, T.; Liu, Z.; Sun, M. JEC-QA: A Legal-Domain Question Answering Dataset. *arXiv* **2019**, arXiv:1911.12011.

11. Carrino, C.P.; Costa-jussà, M.R.; Fonollosa, J.A.R. Automatic Spanish Translation of SQuAD Dataset for Multi-lingual Question Answering. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; European Language Resources Association: Marseille, France, 2020; pp. 5515–5523.
12. d'Hoffschmidt, M.; Belblidia, W.; Heinrich, Q.; Brendlé, T.; Vidal, M. FQuAD: French Question Answering Dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1193–1208. [CrossRef]
13. Abadani, N.; Mozafari, J.; Fatemi, A.; Nematbakhsh, M.A.; Kazemi, A. ParSQuAD: Machine Translated SQuAD dataset for Persian Question Answering. In Proceedings of the 2021 7th International Conference on Web Research (ICWR), Tehran, Iran, 19–20 May 2021; pp. 163–168. [CrossRef]
14. Mozannar, H.; Maamary, E.; El Hajal, K.; Hajj, H. Neural Arabic Question Answering. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, Florence, Italy, 28 July–August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 108–118. [CrossRef]
15. Rogers, A.; Gardner, M.; Augenstein, I. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *arXiv* **2021**, arXiv:2107.12708.
16. van Toledo, C.; van Dijk, F.; Spruit, M. Dutch Named Entity Recognition and De-Identification Methods for the Human Resource Domain. *Int. J. Nat. Lang. Comput.* **2020**, *9*, 23–34. [CrossRef]
17. Kouzis-Loukas, D. *Learning Scrapy*; Packt Publishing Ltd.: Birmingham, UK, 2016.
18. Richardson, L. Beautiful Soup Documentation. April 2007. Available online: https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (accessed on 21 August 2022).
19. Delobelle, P.; Winters, T.; Berendt, B. RobBERT: A Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA 2020; pp. 3255–3265. [CrossRef]
20. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. *spaCy: Industrial-Strength Natural Language Processing in Python*; Zenodo: Honolulu, HI, USA, 2022. [CrossRef]
21. Reeve, J. *Text-Matcher*; GitHub Repository: San Francisco, CA, USA, 2020. [CrossRef]
22. Pander Maat, H.; Kraf, R.; Dekker, N. *Handleiding T-Scan* **2014**. Available online: https://raw.githubusercontent.com/proycon/tscan/master/docs/tscanhandleiding.pdf (accessed on 20 August 2022).
23. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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), Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186. [CrossRef]
24. Xia, P.; Wu, S.; Durme, B.V. Which *BERT? A Survey Organizing Contextualized Encoders. In Proceedings of the EMNLP, Online, 16–20 November 2020.
25. Acheampong, F.A.; Nunoo-Mensah, H.; Chen, W. Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artif. Intell. Rev.* **2021**, *54*, 5789–5829. [CrossRef]
26. de Vries, W.; van Cranenburgh, A.; Bisazza, A.; Caselli, T.; Noord, G.v.; Nissim, M. BERTje: A Dutch BERT Model. *arXiv* **2019**, arXiv:1912.09582.
27. Brandsen, A.; Dirkson, A.; Verberne, S.; Sappelli, M.; Manh Chu, D.; Stoutjesdijk, K. BERT-NL a set of language models pre-trained on the Dutch SoNaR corpus. In Proceedings of the Dutch-Belgian Information Retrieval Conference (DIR 2019), Wuhan, China, 23–27 May 2019.
28. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Ma, C.; Jernite, Y.; Plu, J.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
29. Borzymowski, H. henryk/bert-base-multilingual-cased-finetuned-dutch-squad2 · Hugging Face. In Proceedings of the Benelux Conference on Artificial Intelligence, Esch-sur-Alzette, Luxembourg, 10–12 November 2020.
30. Ohsugi, Y.; Saito, I.; Nishida, K.; Asano, H.; Tomita, J. 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, Florence, Italy, 1 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 11–17. [CrossRef]
31. Qu, C.; Yang, L.; Qiu, M.; Croft, W.B.; Zhang, Y.; Iyyer, M. 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, Paris, France, 21–25 July 2019; ACM: Paris France, 2019; pp. 1133–1136. [CrossRef]