HEAD-QA: A Healthcare Dataset for Complex Reasoning

David Vilares
Universidade da Coruña, CITIC
Departamento de Computación
Campus de Elviña s/n, 15071
A Coruña, Spain
david.vilares@udc.es

Carlos Gómez-Rodríguez
Universidade da Coruña, CITIC
Departamento de Computación
Campus de Elviña s/n, 15071
A Coruña, Spain
carlos.gomez@udc.es

Abstract

We present HEAD-QA, a multi-choice question answering testbed to encourage research on complex reasoning. The questions come from exams to access a specialized position in the Spanish healthcare system, and are challenging even for highly specialized humans. We then consider monolingual (Spanish) and cross-lingual (to English) experiments with information retrieval and neural techniques. We show that: (i) HEAD-QA challenges current methods, and (ii) the results lag well behind human performance, demonstrating its usefulness as a benchmark for future work.

1 Introduction

Recent progress in question answering (QA) has been led by neural models (Seo et al., 2016; Kundu and Ng, 2018), due to their ability to process raw texts. However, some authors (Kaushik and Lipton, 2018; Clark et al., 2018) have discussed the tendency of research to develop datasets and methods that accomodate the data-intensiveness and strengths of *current* neural methods.

This is the case of popular English datasets such as bAbI (Weston et al., 2015) or SQuAD (Rajpurkar et al., 2016, 2018), where some systems achieve near human-level performance (Hu et al., 2018; Xiong et al., 2017) and often surface-level knowledge suffices to answer. To counteract this, Clark et al. (2016) and Clark et al. (2018) have encouraged progress by developing multi-choice datasets that require reasoning. The questions match grade-school science, due to the difficulties to collect specialized questions. With a similar aim, Lai et al. (2017) released 100k questions and 28k passages intended for middle or high school Chinese students, and Zellers et al. (2018) introduced a dataset for common sense reasoning from a spectrum of daily situations.

Question (medicine): A 13-year-old girl is operated on due to Hirschsprung illness at 3 months of age. Which of the following tumors is more likely to be present?
1. Abdominal neuroblastoma
2. Wilms tumor
3. Mesoblastic nephroma
4. Familial thyroid medullary carcinoma.

Question (pharmacology) The antibiotic treatment of choice for Meningitis caused by Haemophilus influenzae serogroup b is:
1. Gentamicin
2. Erythromycin
3. Ciprofloxacin
4. Cefotaxime

Question (psychology) According to research derived from the Eysenck model, there is evidence that extraverts, in comparison with introverts:
1. Perform better in surveillance tasks.
2. Have greater salivary secretion before the lemon juice test.
3. Have a greater need for stimulation.
4. Have less tolerance to pain.

Table 1: Samples from HEAD-QA

However, this kind of dataset is scarce for complex domains like medicine: while challenges have been proposed in such domains, like textual entailment (Abacha et al., 2015; Abacha and Dina, 2016) or answering questions about specific documents and snippets (Nentidis et al., 2018), we know of no resources that require general reasoning on complex domains. The novelty of this work falls in this direction, presenting a multi-choice QA task that combines the need of knowledge and reasoning with complex domains, and which takes humans years of training to answer correctly.

Contribution We present HEAD-QA, a multichoice testbed of graduate-level questions about medicine, nursing, biology, chemistry, psychology, and pharmacology (see Table 1[1]). The data

[1] These examples were translated by humans to English.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 960–966
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

Category	Unsupervised	Supervised setting		
	setting	Train	Dev	Test
Biology	1,132	452	226	454
nursing	1,069	384	230	455
Pharmacology	1,139	457	225	457
Medicine	1149	455	231	463
Psychology	1134	453	226	455
Chemistry	1142	456	228	458
Total	6,765	2,657	1,366	2,742

Table 2: Number of questions in HEAD-QA

Category	Longest question	Avg question	Longest answer	Avg answer
Biology	43	11.11	40	5.08
Nursing	187	29.03	94	9.54
Pharmacology	104	18.18	43	6.70
Medicine	308	55.29	85	9.31
Psychology	103	21.91	43	7.98
Chemistry	63	15.82	52	7.62

Table 3: Tokens statistics in HEAD-QA

is in Spanish, but we also include an English version. We then test models for open-domain and multi-choice QA, showing the complexity of the dataset and its utility to encourage progress in QA. HEAD-QA and models can be found at http://aghie.github.io/head-qa/.

2 The HEAD-QA corpus

The Ministerio de Sanidad, Consumo y Bienestar Social[2] (as a part of the Spanish government) announces every year examinations to apply for specialization positions in its public healthcare areas. The applicants must have a bachelor's degree in the corresponding area (from 4 to 6 years) and they prepare the exam for a period of one year or more, as the vacancies are limited. The exams are used to discriminate among thousands of applicants, who will choose a specialization and location according to their mark (e.g., in medicine, to access a cardiology or gynecology position at a given hospital).

We use these examinations (from 2013 to present) to create HEAD-QA. We consider questions involving the following healthcare areas: medicine (aka MIR), pharmacology (FIR), psychology (PIR), nursing (EIR), biology (BIR), and chemistry (QIR).[3][4] Exams from 2013 and 2014 are multi-choice tests with five options, while the rest of them have just four. The questions mainly refer to technical matters, although some of them also consider social issues (e.g. how to deal with patients in stressful situations). A small percentage (∼14%) of the medicine questions refer to images that provide additional information to answer correctly. These are included as a part of the corpus, although we will not exploit them in this work. For clarity, Table 4 shows an example:[5]

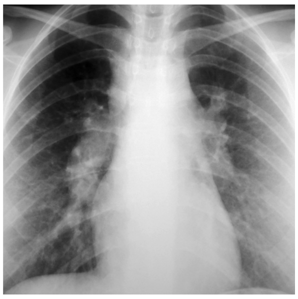

Figure 1: Image no 21 from MIR 2017

Question Question linked to image no 21. A 38-year-old bank employee who has been periodically checked by her company is referred to us to assess the chest X-ray. The patient smokes 20 cigarettes / day from the age of 21. She says that during the last months, she is somewhat more tired than usual. The basic laboratory tests are normal except for an Hb of 11.4 g / dL. An electrocardiogram and forced spirometry are normal. What do you think is the most plausible diagnostic orientation?

1. Hodgkin's disease.
2. Histoplasmosis type fungal infection.
3. Sarcoidosis.
4. Bronchogenic carcinoma.

Table 4: A question referring to Figure 1

We describe in detail the JSON structure of HEAD-QA in Appendix A. We enumerate below the fields for a given sample:

- The question ID and the question's content.
- Path to the image referred to in the question (if any).
- A list with the possible answers. Each answer is composed of the answer ID and its text.
- The ID of the right answer for that question.

[2] https://www.mscbs.gob.es/

[3] Radiophysics exams are excluded, due to the difficulty to parse their content (e.g. equations) from the PDF files.

[4] Some of the questions might be considered invalid after the exams. We remove those questions from the final dataset.

[5] Note that images often correspond to serious injuries and diseases. Viewer discretion is advised. The quality of the images varies widely, but it is good enough that the pictures can be analyzed by humans in a printed version. Figure 1 has 1037x1033 pixels.

Although all the approaches that we will be testing are unsupervised or distant-supervised, we additionally define official training, development and test splits, so future research with supervised approaches can be compared with the work presented here. For this supervised setting, we choose the 2013 and 2014 exams for the training set, 2015 for the development set, and the rest for testing. The statistics are shown in Tables 2 and 3. It is worth noting that a common practice to divide a dataset is to rely on randomized splits to avoid potential biases in the collected data. We decided not to follow this strategy for two reasons. First, the questions and the number of questions per area are designed by a team of healthcare experts who already try to avoid these biases. Second (and more relevant), random splits would impede comparison against official (and aggregated) human results.

Finally, we hope to increase the size of HEAD-QA by including questions from future exams.

English version HEAD-QA is in Spanish, but we include a translation to English (HEAD-QA-EN) using the Google API, which we use to perform cross-lingual experiments. We evaluated the quality of the translation using a sample of 60 random questions and their answers. We relied on two fluent Spanish-English speakers to score the adequacy[6] and on one native English speaker for the fluency,[7] following the scale by Koehn and Monz (2006). The average scores for adequacy were 4.35 and 4.71 out of 5, i.e. most of the meaning is captured; and for fluency 4 out of 5, i.e. good. As a side note, it was observed by the annotators that most names of diseases were successfully translated to English. On the negative side, the translator tended to struggle with elements such as molecular formulae, relatively common in chemistry questions.[8]

3 Methods

Notation We represent HEAD-QA as a list of tuples: $[(q_0, A_0), ..., (q_N, A_N)]$, where: q_i is a question and $A_i = [a_{i0}, ..., a_{im}]$ are the possible answers. We use $\tilde{a}_{ik}$ to denote the predicted answer, ignoring indexes when not needed.

Kaushik and Lipton (2018) discuss on the need of providing rigorous baselines that help better understand the improvement coming from future models, and also the need of avoiding architectural novelty when introducing new datasets. For this reason, our baselines are based on state-of-the-art systems used in open-domain and multi-choice QA (Chen et al., 2017; Kembhavi et al., 2017; Khot et al., 2018; Clark et al., 2018).

3.1 Control methods

Given the complex nature of the task, we include three control methods:

Random Sampling $\tilde{a} \sim Multinomial(\phi)$, where ϕ is a random distribution.

Blind$_x$ $\tilde{a}_{ik} = a_{ix}\ \forall i$. Always chosing the xth option. Tests made by the examiners are not totally random (Poundstone, 2014) and right answers tend occur more in middle options.

Length Choosing the longest answer.[9] Poundstone (2014) points out that examiners have to make sure that the right answer is totally correct, which might take more space.

3.2 Strong multi-choice methods

We evaluate an information retrieval (IR) model for HEAD-QA and cross-lingual models for HEAD-QA-EN. Following Chen et al. (2017), we use Wikipedia as our source of information ($\mathcal{D}$)[10] for all the baselines. We then extract the raw text and remove the elements that add some type of structure (headers, tables, ...).[11]

3.2.1 Spanish information retrieval

Let $(q_i, [a_{i0}, ..., a_{im}])$ be a question with its possible answers, we first create a set of m queries of the form $[q_i + a_{i0}, ..., q_i + a_{im}]$, which will be sent separately to a search engine. In particular, we use the DrQA's Document Retriever (Chen et al., 2017), which scores the relation between the queries and the articles as TF-IDF weighted bag-of-word vectors, and also takes into account word order and bi-gram counting. The predicted answer is defined as $\tilde{a}_{ik} = \arg\max_k(score(m_k, \mathcal{D}))$, i.e.

[6] Adequacy: How much meaning is preserved? We use a scale from 5 to 1: 5 (all meaning), 4 (most meaning), 3 (some meaning), 2 (little meaning), 1 (none).

[7] Fluency: Is the language in the output fluent? We use a scale from 5 to 1: 5 (flawless), 4 (good), 3 (non-native), 2 (disfluent), 1 (incomprehensible).

[8] This particular issue is not only due to the automatic translation process, but also to the difficulty of correctly mapping these elements from PDF exams to plain text.

[9] Computed as the number of characters in the answers.

[10] We downloaded Spanish and English Wikipedia dumps.

[11] github.com/attardi/wikiextractor

the answer in the query m_k for which we obtained the highest document relevance. This is equivalent to the IR baselines by Clark et al. (2016, 2018).

3.2.2 Cross-lingual methods

Although some research on Spanish QA has been done in the last decade (Magnini et al., 2003; Vicedo et al., 2003; Buscaldi and Rosso, 2006; Kamateri et al., 2019), most recent work has been done for English, in part due to the larger availability of resources. On the one hand this is interesting because we hope HEAD-QA will encourage research on multilingual question answering. On the other hand, we want to check how challenging the dataset is for state-of-the-art systems, usually available only for English. To do so, we use HEAD-QA-EN, as the adequacy and the fluency scores of the translation were high.

Cross-lingual Information Retrieval The IR baseline, but applied to HEAD-QA-EN. We also use this baseline as an extrinsic way to evaluate the quality of the translation, expecting to obtain a performance similar to the Spanish IR model.

Multi-choice DrQA (Chen et al., 2017) DrQA first returns the 5 most relevant documents for each question, relying on the information retrieval system described above. It will then try to find the exact span in them containing the right answer on such documents, using a document reader. For this, the authors rely on a neural network system inspired in the Attentive Reader (Hermann et al., 2015) that was trained over SQuAD (Rajpurkar et al., 2016). The original DrQA is intended for open-domain QA, focusing on factoid questions. To adapt it to a multi-choice setup, to select $\tilde{a}$ we compare the selected span against all the answers and select the one that shares the largest percentage of tokens.[12] Non-factoid questions (common in HEAD-QA) are not given any special treatment.

Multi-choice BiDAF (Clark et al., 2018) Similar to the multi-choice DrQA, but using a BiDAF architecture as the document reader (Seo et al., 2016). The way BiDAF is trained is also different: they first trained the reader on SQuAD, but then further tuned to science questions presented in (Clark et al., 2018), using continued training. This system might select as correct more than one answer. If this happens, we follow a simple approach and select the longest one.

Multi-choice DGEM and Decompatt (Clark et al., 2018) The models adapt the DGEM (Parikh et al., 2016) and Decompatt (Khot et al., 2018) entailment systems. They consider a set of hypothesis h_{ik}=$q_i + a_{ik}$ and each h_i is used as a query to retrieve a set of relevant sentences, $\mathcal{S}_{ik}$. Then, an entailment score $entailment(h_{ik}, s)$ is computed for every h_{ik} and $s \in S_{ik}$, where $\tilde{a}$ is the answer inside h_{ik} that maximizes the score. If multiple answers are selected, we choose the longest one.

4 Experiments

Metrics We use accuracy and a POINTS metric (used in the official exams): a right answer counts 3 points and a wrong one subtracts 1 point.[13]

Results (unsupervised setting) Tables 5 and 6 show the accuracy and POINTS scores for both HEAD-QA and HEAD-QA-EN. The cross-lingual IR model obtains even a greater performance than the Spanish one. This is another indicator that the translation is good enough to apply cross-lingual approaches. On the negative side, the approaches based on current neural architectures obtain a lower performance.

	Model	BIR	MIR	EIR	FIR	PIR	QIR	Avg
ES	RANDOM	24.2	22.0	25.1	23.2	24.0	24.5	23.8
	BLIND$_1$	23.7	22.8	22.7	22.4	22.5	21.2	22.5
	BLIND$_2$	25.6	24.3	23.5	23.0	25.3	24.9	24.4
	BLIND$_3$	23.0	24.7	26.5	25.8	22.9	25.1	24.7
	BLIND$_4$	22.6	20.0	21.7	22.4	23.2	22.5	22.1
	LENGTH	26.9	24.9	28.6	28.7	30.6	29.0	28.1
	IR	34.5	26.5	**32.7**	35.5	34.2	**34.2**	32.9
EN	IR	**37.9**	**30.3**	32.6	**38.7**	**34.7**	33.7	**34.6**
	DRQA	29.5	25.0	27.3	28.3	31.0	30.2	28.5
	BIDAF	33.4	26.2	26.8	29.9	26.8	30.3	28.9
	DGEM	31.7	25.7	28.7	29.9	28.5	30.3	29.1
	DECOMPATT	30.6	23.6	27.9	27.2	28.3	27.6	27.5

Table 5: Accuracy on the HEAD-QA and HEAD-QA-EN corpora (unsupervised setting)

Results (supervised setting) We show in Tables 7 and 8 the performance of the top models on the test split corresponding to the supervised setting.

Discussion Medicine questions (MIR) are the hardest ones to answer across the board. We believe this is due to the greater length of both the

[12] We lemmatize and remove the stopwords as in (Clark et al., 2018). We however observed that many of selected spans did not have any word in common with any of the answers. If this happens, we select the longest answer.

[13] Note that as some exams have more choices than others, there is not a direct correspondence between accuracy and POINTS (a given healthcare area might have better accuracy than another one, but worse POINTS score).

	Model	BIR	MIR	EIR	FIR	PIR	QIR	Avg
ES	BLIND3	-17.6	-2.6	16.6	7.4	-18.8	1.2	-2.3
	LENGTH	16.8	-1.0	32.6	33.8	50.8	36.4	28.2
	IR	86.4	14.2	67.0	95.4	82.8	**84.4**	71.7
EN	IR	**116.8**	**48.6**	**67.8**	**125.0**	**87.6**	79.6	**87.6**
	DRQA	40.8	-0.2	20.6	29.8	54.0	47.6	32.1
	BIDAF	75.6	11.0	15.8	44.4	16.6	48.6	35.3
	DGEM	60.8	7.0	34.2	45.0	31.6	48.4	37.8
	DECOMPATT	51.2	-13.0	27.8	20.2	30.0	23.6	23.3

Table 6: POINTS on the HEAD-QA and HEAD-QA-EN corpora (unsupervised setting)

	Model	BIR	MIR	EIR	FIR	PIR	QIR	Avg
ES	RANDOM	24.2	23.1	25.2	23.8	27.9	27.7	25.3
	BLIND3	26.0	27.5	29.8	27.2	24.8	27.8	27.2
	LENGTH	32.4	27.0	32.8	30.2	30.5	30.1	30.5
	IR	36.5	26.3	36.0	40.3	**35.9**	**36.2**	35.2
EN	IR	**39.8**	**33.3**	**36.4**	**42.2**	35.7	36.0	**37.2**
	BIDAF	36.5	26.6	27.7	29.3	28.1	34.1	30.3
	DGEM	31.7	27.2	30.7	29.9	31.0	33.2	30.6

Table 7: Accuracy on the HEAD-QA and HEAD-QA-EN corpora (supervised setting)

questions and the answers (this was shown in Table 3). This hypothesis is supported by the lower results on the nursing domain (EIR), the category with the second longest questions/answers. On the contrary, the categories for which we obtained the better results, such as pharmacology (FIR) or biology (BIR), have shorter questions and answers. While the evaluated models surpass all control methods, their performance is still well behind the human performance. We illustrate this in Table 9, comparing the performance (POINTS score) of our best model against a summary of the results, on the 2016 exams.[14] Also, the best performing model was a non-machine learning model based on standard information retrieval techniques. This reinforces the need for effective information extraction techniques that can be later used to perform complex reasoning with machine learning models.

	Model	BIR	MIR	EIR	FIR	PIR	QIR	Avg
ES	RANDOM	-7.0	-17.5	2.5	-10-5	26.5	25.0	3.2
	BLIND3	9.0	22.5	44.5	19.5	-1.5	25.0	19.8
	LENGTH	67.0	18.5	70.5	47.5	50.5	47.0	50.2
	IR	105.0	12.5	100.5	139.5	**98.5**	**103.0**	93.2
EN	IR	**135.0**	**76.5**	**104.5**	**157.5**	96.5	101.0	**111.8**
	BIDAF	104.0	14.5	18.5	39.0	29.0	83.0	48.0
	DGEM	61.0	20.5	52.5	45.5	54.5	75.0	51.5

Table 8: POINTS on the HEAD-QA and HEAD-QA-EN corpora (supervised setting)

	BIR	MIR	EIR	FIR	PIR	QIR
Avg 10 best humans	627.1	592.2	515.2	575.5	602.1	529.1
Pass mark	219.0	207.0	180.0	201.0	210.0	185.0
EN IR	168.0	124.0	77.0	132.0	62.0	93.0

Table 9: Human performance on the 2016 exams. The results are not strictly comparable, as the last 10 questions are considered as backup questions in the human exams, but still show how far the tested baselines are from human performance.

5 Conclusion

We presented a complex multi-choice dataset containing questions about medicine, nursing, biology, pharmacology, psychology and chemistry. Such questions correspond to examinations to access specialized positions in the Spanish healthcare system, and require specialized knowledge and reasoning to be answered. To check its complexity, we then tested different state-of-the-art models for open-domain and multi-choice questions. We show how they struggle with the challenge, being clearly surpassed by a non-machine learning model based on information retrieval. We hope this work will encourage research on designing more powerful QA systems that can carry out effective information extraction and reasoning.

We also believe there is room for alternative challenges in HEAD-QA. In this work we have used it as a *closed* QA dataset (the potential answers are used as input to determine the right one). Nothing prevents to use the dataset in an *open* setting, where the system is given no clue about the possible answers. This would require to think as well whether widely used metrics such as BLEU (Papineni et al., 2002) or exact match could be appropriate for this particular problem.

Acknowlegments

This work has received support from the TELEPARES-UDC project (FFI2014-51978-C2-2-R) and the ANSWER-ASAP project (TIN2017-85160-C2-1-R) from MINECO, from Xunta de Galicia (ED431B 2017/01), and from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme (FASTPARSE, grant agreement No 714150). We thank Mark Anderson for his help with translation fluency evaluation.

[14] 2016 was the annual examination for which we were able to find more available information.

References

Asma Ben Abacha and Demner-Fushman Dina. 2016. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium Proceedings*, volume 2016, page 310. American Medical Informatics Association.

Asma Ben Abacha, Duy Dinh, and Yassine Mrabet. 2015. Semantic analysis and automatic corpus construction for entailment recognition in medical texts. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 238–242. Springer.

Davide Buscaldi and Paolo Rosso. 2006. Mining knowledge from wikipedia for the question answering task. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 727–730.

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.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *AAAI*, pages 2580–2586.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. Reinforced mnemonic reader for machine reading comprehension. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*.

Eleni Kamateri, Theodora Tsikrika, Spyridon Symeonidis, Stefanos Vrochidis, Wolfgang Minker, and Yiannis Kompatsiaris. 2019. A test collection for passage retrieval evaluation of spanish health-related resources. In *European Conference on Information Retrieval*, pages 148–154. Springer.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4999–5007.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of AAAI*.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Souvik Kundu and Hwee Tou Ng. 2018. A nil-aware answer extraction framework for question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4243–4252, Brussels, Belgium. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Bernardo Magnini, Simone Romagnoli, Alessandro Vallin, Jesús Herrera, Anselmo Penas, Víctor Peinado, Felisa Verdejo, and Maarten de Rijke. 2003. The multiple language question answering track at clef 2003. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 471–486. Springer.

Anastasios Nentidis, Anastasia Krithara, Konstantinos Bougiatiotis, Georgios Paliouras, and Ioannis Kakadiaris. 2018. Results of the sixth edition of the bioasq challenge. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 1–10. Association for Computational Linguistics.

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

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255. Association for Computational Linguistics.

William Poundstone. 2014. *Rock breaks scissors: a practical guide to outguessing and outwitting almost everybody*. Hachette UK.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. pages 784–789.

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.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

José L Vicedo, Ruben Izquierdo, Fernando Llopis, and Rafael Munoz. 2003. Question answering in spanish. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 541–548. Springer.

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*.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dcn+: Mixed objective and deep residual coattention for question answering. *arXiv preprint arXiv:1711.00106*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

A Appendices

We below describe the `fields` of the JSON file use to represent HEAD-QA.

```
{
 "version": 1.0
 "language": ["es","en"]
 "exams": A list of exams.
    "name": Cuaderno_YEAR_1_*IR_ACRONYM.
    "year": e.g. 2016.
    "category": ['medicine','biology',
                 'nursing','pharmacology',
                 'chemistry','psychology']
    "data": A list of questions/answers.
       "qid": The question ID, extracted
              from the original PDF exam
              (usually between 1 and 235).
       "qtext" : The text of the question.
       "ra" : The ID of the right answer.
       "answers": A list with the answer options.
          "aid": The answer ID (1 to 5).
          "atext": The text of the answer.
}
```