CliCR: A Dataset of Clinical Case Reports for Machine Reading Comprehension*

Simon Šuster and **Walter Daelemans**
Computational Linguistics & Psycholinguistics Research Center,
University of Antwerp, Belgium
{simon.suster,walter.daelemans}@uantwerpen.be

Abstract

We present a new dataset for machine comprehension in the medical domain. Our dataset uses clinical case reports with around 100,000 gap-filling queries about these cases. We apply several baselines and state-of-the-art neural readers to the dataset, and observe a considerable gap in performance (20% F1) between the best human and machine readers. We analyze the skills required for successful answering and show how reader performance varies depending on the applicable skills. We find that inferences using domain knowledge and object tracking are the most frequently required skills, and that recognizing omitted information and spatio-temporal reasoning are the most difficult for the machines.

1 Introduction

Machine comprehension is a task in which a system reads a text passage and then answers questions about it. The progress in machine comprehension heavily depends on the introduction of new datasets (Burges, 2013), which encourages the development of new algorithms and deepens our understanding of the (linguistic) challenges that can or can not be tackled well by these algorithms. Recently, a number of reading comprehension datasets have been proposed (§ 2), differing in various aspects such as mode of construction, answer-query formulation and required understanding skills. Most are open-domain datasets built from news, fiction and Wikipedia texts. For specialized domains, however, large machine comprehension datasets are extremely scarce (Welbl et al., 2017a), and the required comprehension skills poorly understood. With our work we hope to narrow this gap by proposing a new resource for reading comprehension in the clinical domain, and by analyzing the different types of comprehension skills that are triggered while answering (Sugawara et al., 2017; Lai et al., 2017).

passage:
[...] A gradual improvement in clinical and laboratory status was achieved within 20 days of antituberculous treatment . The patient was then subjected to a thoracic CT scan that also showed significant radiological improvement . *Thereafter , tapering of corticosteroids was initiated with no clinical relapse* . The patient was discharged after being treated for a total of 30 days and continued receiving antituberculous therapy with no reported problems for a total of 6 months under the supervision of his hometown physicians . [...]
query:
If steroids are used , great caution should be exercised on their gradual tapering to avoid _____ .
answer:
relapse (sem_type=problem, cui=C0035020)

Figure 1: An example from the dataset, with the passage sentence relevant for answering italicized. The passage has been shortened for clarity.

Machine comprehension for healthcare and medicine has received little attention so far, although it offers great potential for practical use. A typical application would be clinical decision support, where given a massive amount of text, a clinician asks questions about either external, medical knowledge (reading literature) or about particular patients (reading electronic health records). Currently, patient-specific questions are tackled by manually browsing or searching those records. This task can be facilitated by summarization and QA systems (Demner-Fushman and Lin, 2007; Demner-Fushman et al., 2009), and we believe, by fine-grained machine reading. Reading comprehension systems that perform on a finer level could play an important role especially when combined with

*We provide the information about accessing the dataset, as well as the code for the experiments, at http://github.com/clips/clicr.

Proceedings of NAACL-HLT 2018, pages 1551–1563
New Orleans, Louisiana, June 1 - 6, 2018.

document retrieval to perform machine reading at scale, such as in the models of Chen et al. (2017) and Watanabe et al. (2017) for the general domain.

For our dataset, we construct queries, answers and supporting passages from BMJ Case Reports, the largest online repository of such documents. A case report is a detailed description of a clinical case that focuses on rare diseases, unusual presentation of common conditions and novel treatment methods. Each report contains a *Learning points* section, summarizing the key pieces of information from that report. The learning points are typically paraphrased portions of passage text and do not match passage sentences exactly. We use these learning points to create queries by blanking out a medical entity. To counteract potential errors and inconsistencies due to automated dataset creation, we perform several checks to improve the quality of the dataset (§ 3). Our dataset contains around 100,000 queries on 12,000 case reports, has long support passages (around 1,500 tokens on average) and includes answers which are single- or multi-word medical entities. We show an example from the dataset in Figure 1.

We examine the performance on the dataset in two ways. First, we report machine performance for several baselines and neural readers. To enable a more flexible answer evaluation, we expand the answers with their respective synonyms from a medical knowledge base, and additionally supplement the standard evaluation metrics with BLEU and embedding-based methods. We investigate different ways of representing medical entities in the text and how this affects the neural readers. We obtain the best results with a recurrent neural network (RNN) with gated attention (Dhingra et al., 2017a), but a simple approach based on embedding similarity proves to be a strong baseline as well. Second, we look at how well humans perform on this task, by asking both a medical expert and a novice to answer a portion of the validation set. When categorizing the skills necessary to find the right answer, we observe that a large number of comprehension skills get activated and that prior knowledge in the form of the ability to perform lexico-grammatical inferences matters the most. This suggests that for our dataset and possibly for domain-specific datasets more generally, more background knowledge should be incorporated in machine comprehension models. The current gap between the best machine and the best human performance is nearly 20% F1, which leaves ample space for further study of machine readers on our dataset. In brief, the contributions of our paper are:

- We propose a large dataset for reading comprehension in the medical domain, using clinical case descriptions.
- We carry out an empirical analysis of *a*) system and human performance on reading comprehension, and *b*) comprehension skills that are required for answering the queries correctly and that allow us to position the dataset according to its difficulty on each of the skills.

Dataset	Question origin	Domain	Size
CliCR (this work)	Learning points	Medical	105K
Quasar-S (Dhingra et al., 2017b)	Definitions	Software	37K
SciQ (Welbl et al., 2017a)	Crowdsourced	Science	14K
MedHop (Welbl et al., 2017b)	KB	Drugs	2.5K
Biology (Berant et al., 2014)	Domain expert	Biology	585
Algebra (Kushman et al., 2014)	Crowdsourced	Algebra	514
QA4MRE (Sutcliffe et al., 2013)	Annotator	Various	240

Table 1: Survey of closed-domain reading comprehension datasets. Size: number of questions. We did not include remotely related datasets which concern a different task (e.g. information retrieval) (Roberts et al., 2015; Voorhees and Tice, 2000).

2 Related datasets

Numerous **general-domain datasets** have been recently created to allow machine comprehension using data-intensive methods. These datasets were collected from Wikipedia (Hewlett et al., 2016; Joshi et al., 2017; Rajpurkar et al., 2016), web search queries (Nguyen et al., 2016), news articles (Hermann et al., 2015; Onishi et al., 2016; Trischler et al., 2017), books (Bajgar et al., 2016; Hill et al., 2016; Paperno et al., 2016) and English exams (Lai et al., 2017). In Table 1, we compare our dataset to several **domain-specific datasets** for machine comprehension. In Quasar-S, the queries are constructed from definitions of software entity tags in a community QA website, while in our case the queries are more varied and explicitly relate to the supporting passages. SciQ is a dataset of science exam questions, in which question-answer pairs are used to retrieve the text passages. For each question, four candidate answers are available. In our dataset, the number of candidate answer is much

higher as the candidate answers come from the relatively long passages. Other datasets mentioned in the table are smaller, so they could not be used as training sets for statistical NLP models.

Cloze datasets require the reader to fill in gaps by relying on accompanying text. Representative datasets are Children's Book Test (Hill et al., 2016) and Book Test (Bajgar et al., 2016), in which queries are created by removing a word or a named entity from the running text in a book; and Hermann et al. (2015), who similarly to us blank out entities in abstractive CNN and Daily Mail summaries, but who are only concerned with short proper nouns and short passages. Who-did-what (Onishi et al., 2016) requires the reader to select the person name from a short candidate list that best answers the query about a news event. They do not use summaries for query formation but remove a named entity from the initial sentence in a news article, and then perform information retrieval to find independent passages relevant to the query. Another cloze dataset for language understanding is ROCStories (Mostafazadeh et al., 2016), but it is targeted more towards script knowledge evaluation, and only contains five-sentence stories. Another related task is predicting rare entities only, with a focus on improving a reading comprehension system with external knowledge sources (Long et al., 2017).

Another popular way of creating datasets for reading comprehension is **crowdsourcing** (Rajpurkar et al., 2016; Richardson et al., 2013; Nguyen et al., 2016; Trischler et al., 2017). These datasets exist primarily for the general domain; for specialized domains where background knowledge is crucial, crowdsourcing is intuitively less suitable (Welbl et al., 2017b), although some positive precedent exists for example in crowdsourcing annotations of radiology reports (Cocos et al., 2015). Compared to automated dataset construction, crowdsourcing is more likely to provide high-quality queries and answers. On the other hand, human question generation may also lead to less varied datasets as questions would tend to be of *wh-* type; for cloze datasets, the questions may be more varied and might require readers to possess a different set of skills.[1]

[1]Support for this is given in Sugawara et al. (2017), who show that Who-did-what dataset, for example, requires on average a larger number of reading skills than SQuAD (Rajpurkar et al., 2016) and MCTest (Richardson et al., 2013).

3 Dataset design

We collected the articles from BMJ Case Reports[2]. The data span the years 2005–2016 and amount to almost 12 thousand reports. We removed the HTML boilerplate from the crawled reports using jusText[3], segmented and tokenized the texts with cTakes (Savova et al., 2010), and annotated the medical entities using Clamp (Soysal et al., 2017). We apply two simple heuristics to refine the recognized entities and to decrease their sparsity. Namely, we move the function words (determiners and pronouns) from the beginning of the entity outside of it, and we adjust the entity boundary so that it does not include a parenthetical at the end of the entity. Clamp assigns entities following the i2b2-2010 shared task specifications (Uzuner et al., 2011). For each entity, a concept unique identifier (CUI) is also available, which links it to the UMLS® Metathesaurus® (Lindberg et al., 1993). To check the quality of the recognized entities, we carried out a small manual analysis on 250 entities. We found that in 89% of cases, the boundaries were correct and defined a true entity. Wrongly recognized cases occurred mostly when two entities were coordinated and recognized as one; when a verb was wrongly included in the entity; or when a pre-modifier was left out.

3.1 Query construction

We create a query by replacing a medical entity in one learning point with a blank. For example, in a report describing comorbid disorders of ADHD, we could obtain the following query:

(1) "Patients with ADHD have higher incidence of ___."

The missing entity "enuresis" is taken as the correct answer. Even though one query corresponds to at most one learning point, there can be more than one query built from a learning point. Occasionally, a learning point contains an exact repetition from the passage. These instances would be trivial to answer, so we remove them. We count as an exact match every instance whose longer side to left/right of the query blank coincides with a part in the passage text. This curation step reduces the dataset size by 5%. More commonly, the learning points are paraphrases of crucial parts of the passage. Sometimes, the entity answering the query is expressed

[2]http://casereports.bmj.com/
[3]https://pypi.python.org/pypi/jusText

differently in the passage. For example, in place of "enuresis", the passage might include its synonym "bedwetting". We manage these cases in two ways, by extending the set of answers for a certain query (§ 3.2), and adding a semantic relatedness metric to the standard evaluation (§ 6).

3.2 Answer set

We account for lexical variation of the ground-truth answers (compared to mentions in the passages) by extending each original ground-truth answer a to a set of ground-truth answers A using a knowledge base. Since our entity recognizer already provides the CUI labels, we can use them to obtain the list of alternative word and phrase forms (synonyms, abbreviations and acronyms) from UMLS®.

Similarly to previous work (Choi et al., 2016; Hewlett et al., 2016), for certain queries none of the answers in A occurs verbatim in the passage. We have found upon manual inspection that this is mostly due to lexical variation that is not captured by answer extension, and to a lesser degree, due to the introduction of entirely new information in the learning point and the entity recognition errors. In the empirical part, we use for training only the instances for which at least one answer occurs in the passage, but we evaluate on all instances in the validation and test sets, including those for which $A \cap E = \emptyset$, where E is the set of all entities in the passage. This mimics a likely real-life scenario where the set of ground-truth answers is a priori unknown.

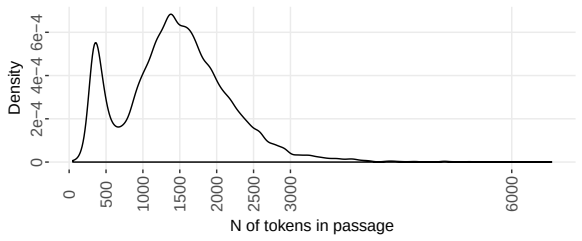
3.3 Task formulation

The reading comprehension problem in our case can be represented as a tuple (q, p, A), where q is the query, built from a learning point; the passage p is the entire report excluding the Learning points section; and A is the set of ground-truth entities answering q. In defining the task, it is important to consider how to take into account entity annotation and how to define the answer output space. We look at these more closely in the rest of this section.

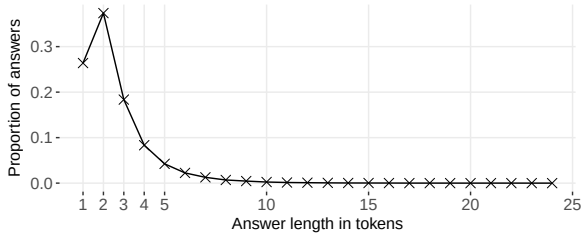
Whenever the entities are marked in the passage, the system can learn to exploit this cue to find the answers more easily (Wang et al., 2017). Although this simplifies the task, it also makes it less realistic as the entities may not be recognized at test time. Realizing that the presence of entities makes the task easier for the machines, Hermann et al. (2015) anonymize the entities, also with a goal of discouraging language model solutions to the queries. In our case, it is not clear how relevant the anonymization is since we deal with medical entities, which have different properties than proper name entities (Kim et al., 2003; Niu et al., 2003). We explore different entity-annotation choices in the empirical part, where we refer to them as **Ent** (entities marked) and **Anonym** (entities marked but anonymized). We further examine a more challenging setup in which the reader can not rely on entity markers as they are not present in the passage (**NoEnt**). In all cases, the reader chooses an answer among the candidates E collected from all entities in the passage.[4] Multi-word entities, which are common in our dataset, are treated as a single token by Ent and Anonym.

N of cases	11,846
N of queries in train/dev/test	91,344/6,391/7,184
N of tokens in passages	16,544,217
N of word types in passages	112,673
N of entity types in passages	591,960
N of distinct answers	56,093
N of distinct answers (incl. extended)	288,211
% answers verbatim in passage	59

Table 2: Data statistics based on the lowercased dataset. For *N of tokens in passages*, we count each passage exactly once, although several queries are normally associated with a passage.



(a)



(b)

Figure 2: Distribution of (a) passage and (b) answer length. Curve (a) is bimodal due to shorter lengths of articles published prior to 2008.

[4] The candidate answers could in principle be obtained also in some other way, so we do not list them in our dataset.

Type	%	Example
problem	67	tuberculosis, abdominal pain, acute myocardial infarction
treatment	22	chemotherapy, surgical intervention, vitamin D suppl.
test	11	MRI, histopathological exam.

Table 3: Answer type statistics.

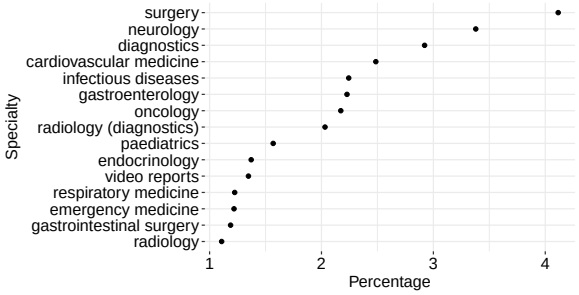


Figure 3: The 15 most common medical specialties represented in the dataset.

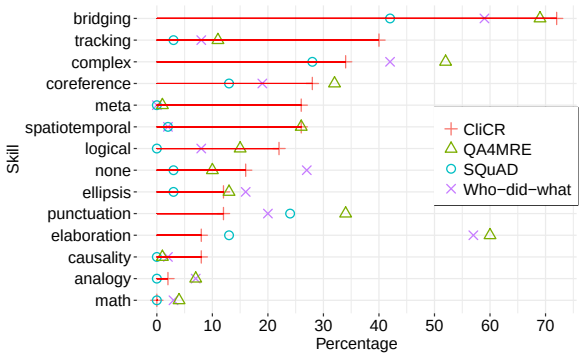


Figure 4: Percentage of times a skill is required in a given dataset. The percentages for the datasets other than ours are from Sugawara et al. (2017).

4 Dataset analysis

We now describe the dataset in more detail, starting with the general statistics summarized in Table 2. It is worth pointing out that the support passages are rather long, which stems from the data origin (journal articles). We show the passage length distribution in Figure 2a, which has the average length of 1,466 tokens. Furthermore, passages are rich with medical entities. There is little repetition of answers—the total of around 100,000 queries are answered by 50,000 distinct entities. Upon extending the answer set with UMLS® we introduce on average four alternative answers for each original one. In 59% of instances, the answer entity is found verbatim in the relevant passage. The answers can belong to any of the problem, treatment or test categories (Table 3), and usually consist of multiple words (Figure 2b). The diversity of medical specialties represented in the articles is shown in Figure 3.

4.1 Analysis of comprehension skills

We estimate the types of skills required in answering by following the categorization of Sugawara et al. (2017). We include the skill definitions with examples from our dataset in Appendix B. We annotated 100 instances in the validation set (with ground-truth answers provided), which yielded on average 2.85 skills per query. The distribution of the required skills is shown in Figure 4. In comparison to the general-domain datasets (SQuAD, Who-did-what), our dataset and QA4MRE (which is also a domain-specific dataset, but with human-generated questions) require more bridging inferences (inferences using background knowledge about the domain), spatio-temporal reasoning and coreference resolution. In our dataset, meta knowledge and object tracking are required more often than in any other dataset. This can be explained by the data origin and the nature of queries. In the case reports, a prominent topic can be discussed which the author refers to in the query, but the query itself is never answered in the passage (meta knowledge). Furthermore, the authors often enumerate medical entities in the query, which leads to the frequent use of object tracking. The queries which were unanswerable are marked as "none". The fraction of these cases was around 16%.

In our experience, the annotation of skills proved quite challenging due to certain confusables. For example, object tracking and coreference both need to maintain the link between objects; object tracking, which includes establishing set relations and membership, may be overlaid with the schematic clause relation skill (subordination); and bridging inference can overlap with coreference resolution. Nevertheless, we adhered to this classification of skills to increase comparability to other datasets included in Figure 4.

5 Methods

5.1 Baselines

Our simplest baselines that we apply on the test set include choosing a random entity (**rand-entity**)

and selecting the most frequent passage entity (**maxfreq-entity**) as the answer. We also include a distance-based method that uses word embeddings (**sim-entity**). Here, we vectorize the passage and the query, and then choose that entity from the passage whose representation has the highest cosine similarity to the query representation:

$$\textit{sim-entity} = \underset{i \in E}{\operatorname{argmax}} \cos\Big(\sum_{j \in C_i} c_j, \sum_{k \in Q} q_k \Big), \quad (1)$$

where $c, q \in \mathbb{R}^d$. The multiset C_i contains the words $\{x_{i-n}, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{i+n}\}$ surrounding the passage entity $i \in E$. We define Q, the context words of the query, likewise. To find out how well the queries can be answered without reading the passage, we also predict the most likely continuation with a language model (**lang-model**). We trained a 4-gram Kneser-Ney model on CliCR training data (with multi-word entities represented as a single token) using SRILM (Stolcke, 2002).

5.2 Neural readers

We apply two types of bidirectional RNNs to our data. Following Wang et al. (2017), we distinguish between *aggregation* readers and *explicit reference* readers, which differ in their formulation of the attention mechanism and how it is being used for answer prediction.

Stanford Attentive (SA) Reader The model proposed by Chen et al. (2016) is an aggregation reader based on the Attentive Reader (Hermann et al., 2015). It predicts the answer using:

$$\hat{a} = \underset{i \in E}{\operatorname{argmax}}\, e_o(i)^T o, \quad (2)$$

where $e_o(i)$ is the answer's output embedding and o is the passage representation obtained by weighting every token representation in the passage with attention: $o = \sum_t \alpha_t h_t$. The attention mechanism is used here to measure the compatibility between token (h_t) and query (q) representations with a bilinear form, $\alpha_t = \text{softmax}_t h_t^T W_\alpha q$. At prediction time, attention should highlight that position t in the passage where the answer occurs. Note that the prediction relies on the aggregate representation o, hence the name of the reader category. As we see in (2), the prediction score does not allow accounting for multi-word entities, unless they are treated as a single token. Returning to our different set-ups based on entity annotation (§ 3.3), this means that we can apply SA reader with Ent and Anonym set-ups, but not with NoEnt, where multi-word answers should be allowed.

Gated-Attention (GA) Reader Dhingra et al. (2017a) investigate neural readers with a fine-grained attention mechanism that learns token representations for the passage that are also conditional on the query, but are in addition refined through multiple hops of the network. The model predicts the answer using attention weights with *explicit reference* to answer positions in the passage:

$$\hat{a} = \underset{i \in E}{\operatorname{argmax}} \sum_{t \in R(i,p)} \alpha_t, \quad (3)$$

where R is the set of indices in passage p at which a token from the candidate i occurs. This operation is also called the pointer sum attention (Kadlec et al., 2016). Since the model marks the references for each token in the answer separately, it allows us to investigate also the NoEnt set-up.[5]

We train each reader with the best hyper-parameters found on the validation set using random search (Bergstra and Bengio, 2012), and evaluate it on the test part of the dataset. We provide more details about parameter optimization in Appendix A. The models use word embeddings pre-trained on biomedical texts.

5.3 Embedding data and pre-training

We induce the word embeddings on a combination of the CliCR training corpus and PubMed abstracts with open-access PMC articles available until 2015 (segmented and tokenized), amounting to over 9 billion tokens (Hakala et al., 2016). Considering the large effect of hyper-parameter selection on the quality of word embeddings (Levy et al., 2015), we optimize the embedding hyper-parameters also using random search.

6 Evaluation

A model f takes as input a passage–query pair and outputs an answer $\hat{a}$.[6] We carry out the evaluation

[5] We assume the candidate entities are known in advance.

[6] In our case, the answer is a word or a word phrase representing a medical entity. Alternatively, one could also take the UMLS® CUI identifier as the answering unit. However, in that case, it would mean that sometimes the original word phrase is lost. This is because entity linking with CUIs can be noisy, and only a part of a word phrase may be linked to the ontology. In the current setup, we are able to keep both the original word phrase as well as the extended answers. The CUI information is still an integral part of the answer field in our dataset, so it can be used by other researchers if preferred.

with different metrics described below. The final score m for a metric v is obtained by averaging over the test set:

$$m_v(f) = \frac{1}{|D_{\text{test}}|} \sum_{(p,q,A) \in D_{\text{test}}} \max_{a \in A} v(f(p,q), a). \quad (4)$$

Since there are multiple correct answers A, we take the highest scoring answer $\hat{a}$ at each instance, as done in Rajpurkar et al. (2016). Note that in the dataset we do not supply the candidate answers; in the experiments, we constrain the candidates to the set of entities in the passage.

The two standardly used metrics for machine comprehension evaluation are the **exact match** (EM) and the **F1** score. For EM, the predicted and the ground truth answers must match precisely, safe for articles, punctuation and case distinction (same for other metrics). F1 metric is applied per instance and measures the overlap between the prediction $\hat{a}$ and the ground truth a, which are treated as bags of words.[7] While these two metrics are arguably sufficient in news-style machine comprehension where the entities are proper nouns which allow for little variation and synonymy, in our case the medical entities are often mostly common nouns modified by specifiers and qualifiers. To take into account potentially large lexical and word-order variation, we use two additional metrics. First, we measure **BLEU** (Papineni et al., 2002) for n-grams of length 2 (shortly, B2) and 4 (B4) using the package by Chen et al. (2015), with which we aim to capture contiguity of tokens in longer answers. Second, it may occur that answers contain no word overlap yet still be good candidates because of their semantical relatedness, as in "renal failure"–"kidney breakdown". We take this into account by using an **embedding metric** (Emb), in which we construct mean vectors for both ground-truth and system answer sequences, and then compare them with the cosine similarity. This and other embedding metrics for evaluation were previously studied in dialog-system research (Liu et al., 2016).

7 Results and analysis

We show the results in Table 4. We see that answer prediction based on contextual representation of queries and passages (sim-entity) achieves a strong base performance that is only outperformed by GA reader. The language model performs poorly on EM and F1, but the embedding-metric score is higher, likely reflecting the fact that the predicted answers—though mostly incorrect—are related to the ground-truth answers. The poor performance means that based on queries alone (without reading the passage), it is difficult to provide accurate answers. The GA reader performs well across all entity set-ups, even when the entities are not marked in the passage. Interestingly, the exact match and BLEU scores in this case are much lower compared to other entity set-ups. Upon inspecting the predicted answers more closely, we have observed that GA-NoEnt tends to predict longer answers than GA-Ent/Anonym. For example, the average predicted answer length for GA-NoEnt was as high as 3.7 tokens, whereas for the other two set-ups and the ground-truth answers the numbers range between 2.3 and 2.5. A plausible explanation for this lies in how GA reaches its prediction (3), which is by accumulating the attention weights without normalizing. This would then drive the model to prefer longer answers. For example, for the ground-truth entity "chest CT", GA-NoEnt predicts "interval CT scans of the chest". Although all neural models use pre-trained word embeddings, for Ent and Anonym the multi-word entities do not have pre-trained embeddings since our embeddings are induced on the word level. This may partly explain the competitive performance of NoEnt compared to Ent. We leave the integration of entity embeddings for the future work.

The results for SA reader are far below the per-

Method	EM	F1	B2	B4	Emb
rand-entity	1.4	5.1	.03	.01	.23
maxfreq-ent.	8.5	12.6	.10	.05	.31
sim-entity	20.8	29.4	.22	.15	.45
lang-model	2.1	3.5	.00	.00	.30
SA-Anonym	19.6	27.2	.22	.16	.43
SA-Ent	6.1	11.4	.07	.05	.31
GA-Anonym	24.5	33.2	.28	.20	.48
GA-Ent	22.2	30.2	.25	.18	.46
GA-NoEnt	14.9	33.9	.21	.11	.51
human-expert	*35*	*53.7*	*.46*	*.23*	*.67*
human-novice	*31*	*45.1*	*.43*	*.24*	*.62*

Table 4: Answering results on the test set. EM and F1 scores are percentages. The human scores (in italics) are based on the validation set.

[7] In precision, the number of correct words is divided by the number of all predicted words. In recall, the former is divided by the number of words in the ground-truth answer.

formance of GA reader. We also see that it performs much better on anonymized entities than on non-anonymized ones. This is in line with Wang et al. (2017) who find that SA reader suffers a drop of 19 points in exact match on Who-did-what dataset when anonymization is not done. A possible explanation is that anonymization reduces the output space to only several hundred entity candidates for which the output embedding needs to be trained. When we do not use anonymization, the set of output entities increases to the set of all entity types found in all passages, which is several orders of magnitude more. While this effect also occurs for GA reader, it is less pronounced because GA reader scores words in the passage and does not need to learn separate answer word embeddings.

7.1 Human performance

To measure the accuracy of human answering, we have used the same sample of data instances as used for the analysis of skills.[8] The queries were answered separately by a novice reader (linguistics background, little-to-none medical knowledge) and by an expert reader (both linguistics and medical background). The annotators needed around 15 minutes on average to read the passage and answer the query. The results are shown at the bottom of Table 4. The expert scores higher across all evaluation metrics, with as much as a 7-point advantage in % F1. This advantage is largely coming from the better performance on those instances where bridging inferences are required (the average F1 score was 10 points higher on these queries), which suggests that domain knowledge is beneficial in the comprehension task. For a novice in a specialized domain, it is harder to build a good situation model that would lead to successful comprehension since it requires more effort—active, strategic processing and establishing ontological relationships in that specific domain. For an expert reader this process is more automatized (Kintsch and Rawson, 2008).

We can see from the table that the best human performance is well below its theoretical upper bound of 100% F1. An important part of explanation for this lies in the automated dataset construction, which leaves certain queries unanswerable, especially when the authors do not refer to a part in the article but introduce completely new information. Another reason is the problem of "answer openness": Typically more than one correct answer is possible and the answers can be correct to various degrees, which we aimed to capture with the use of the embedding metric in the evaluation. Nevertheless, the gap between the best human and machine F1 score is large (around 20 points), leaving considerable space for future applications of machine readers on our dataset.[9]

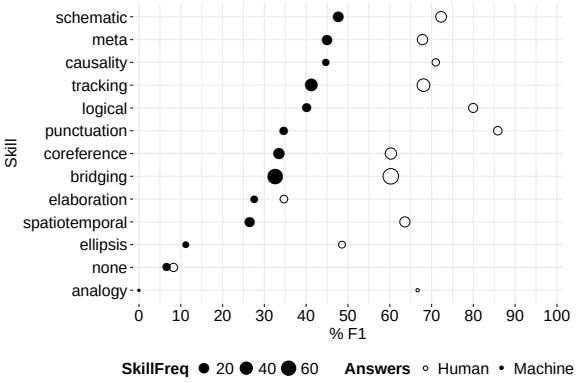


Figure 5: Performance per required skill for the human expert and GA-NoEnt reader.

7.2 Breakdown of results by skill

To see how the answering performance relates to the skill requirements, we have analyzed the part of the validation set annotated with the skills by averaging F1 values for all instances with a particular skill. In this way, we are able to break down both human and machine performance skill-wise, as shown in Figure 5. Because of the small sample size, the results should only be taken as a general indication. The most difficult cases for the GA reader are those annotated with "none" (unanswerable) and "ellipsis" (recognizing implicit and omitted information), ignoring "analogy" for which we only have a single annotated case. Furthermore, spatio-temporal reasoning, elaboration (inferences using general knowledge) and bridging—which is also the most commonly required skill—are the next most difficult ones. The human scores are mostly much higher, which is especially apparent for spatio-temporal reasoning, logical skills and the skill involving punctuation. Our findings align with those of Chu et al. (2017) on the Lambada dataset (Paperno et al., 2016): Although they used a different categorization of comprehension skills, they also find that GA reader has most difficulties with elaboration (which they refer to as "external

[8] Human answers were collected before the skill analysis.

[9] For comparison, the gap for SQuAD was 12.2 and for NewsQA 19.8 (Trischler et al., 2017).

knowledge"), followed by coreference resolution.

8 Conclusion and future work

We have introduced a new dataset for domain-specific reading comprehension in which we have constructed around 100,000 cloze queries from clinical case reports. We analyzed the dataset in terms of the skills required for successful comprehension, and applied various baseline methods and state-of-the-art neural readers. We showed that a large gap still exists between the best machine reader and the expert human reader. One direction for future research is improving the reading models on the queries that are currently the most challenging, i.e. those requiring world and background domain knowledge. Better representing background knowledge by inducing embeddings for entities or otherwise integrating ontological knowledge is in our opinion a promising avenue for future research.

Acknowledgments

We would like to thank Madhumita Sushil and the anonymous reviewers for useful comments. We are also grateful to BMJ Case Reports for allowing the collection of case reports. This work was carried out in the framework of the Accumulate IWT SBO project (nr. 150056), funded by the government agency for Innovation by Science and Technology. We also acknowledge the support of the Nvidia GPU Grant Program.

References

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2016. Embracing data abundance: Booktest dataset for reading comprehension. *arXiv preprint arXiv:1610.00956* .

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 1499–1510. https://doi.org/10.3115/v1/D14-1159.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13(Feb):281–305.

Christopher JC Burges. 2013. Towards the machine comprehension of text: An essay. Technical report, Microsoft Research Technical Report MSR-TR-2013-125, 2013, pdf.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 2358–2367. https://doi.org/10.18653/v1/P16-1223.

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)*. Association for Computational Linguistics, pages 1870–1879. https://doi.org/10.18653/v1/P17-1171.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* .

E. Choi, D. Hewlett, A. Lacoste, I. Polosukhin, J. Uszkoreit, and J. Berant. 2016. Hierarchical question answering for long documents. *arXiv preprint arXiv:1611.01839* .

Zewei Chu, Hai Wang, Kevin Gimpel, and David McAllester. 2017. Broad context language modeling as reading comprehension. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, pages 52–57. http://www.aclweb.org/anthology/E17-2009.

Anne Cocos, Aaron Masino, Ting Qian, Ellie Pavlick, and Chris Callison-Burch. 2015. Effectively crowdsourcing radiology report annotations. In *Sixth International Workshop on Health Text Mining and Information Analysis (Louhi)*. https://doi.org/10.18653/v1/W15-2614.

Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics* 42(5):760 – 772.

Dina Demner-Fushman and Jimmy Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics* 33(1):63–103.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017a. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1832–1846. https://doi.org/10.18653/v1/P17-1168.

Bhuwan Dhingra, Kathryn Mazaitis, and William W Cohen. 2017b. Quasar: Datasets for Question Answering by Search and Reading. *arXiv preprint arXiv:1707.03904* .

Kai Hakala, Suwisa Kaewphan, Tapio Salakoski, and Filip Ginter. 2016. Syntactic analyses and named entity recognition for pubmed and pubmed central — up-to-the-minute. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, pages 102–107. https://doi.org/10.18653/v1/W16-2913.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. WikiReading: A Novel Large-scale Language Understanding Task over Wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1535–1545. https://doi.org/10.18653/v1/P16-1145.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. In *ICLR*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1601–1611. https://doi.org/10.18653/v1/P17-1147.

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 908–918. https://doi.org/10.18653/v1/P16-1086.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. GENIA corpus: a semantically annotated corpus for bio-textmining. *Bioinformatics* 19(suppl 1):i180–i182.

Walter Kintsch and Katherine A. Rawson. 2008. *Comprehension*, Blackwell Publishing Ltd, pages 211–226. https://doi.org/10.1002/9780470757642.ch12.

Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and Regina Barzilay. 2014. Learning to automatically solve algebra word problems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 271–281. https://doi.org/10.3115/v1/P14-1026.

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*. Association for Computational Linguistics, pages 796–805. http://aclweb.org/anthology/D17-1083.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association of Computational Linguistics* 3:211–225. http://www.aclweb.org/anthology/Q15-1016.

Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The Unified Medical Language System. *Methods of information in medicine* 32(04):281–291.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2122–2132. http://aclweb.org/anthology/D16-1230.

Teng Long, Emmanuel Bengio, Ryan Lowe, Jackie Chi Kit Cheung, and Doina Precup. 2017. World knowledge for reading comprehension: Rare entity prediction with hierarchical lstms using external descriptions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 825–834. http://aclweb.org/anthology/D17-1086.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR Workshop Papers*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 839–849. http://www.aclweb.org/anthology/N16-1098.

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

Yun Niu, Graeme Hirst, Gregory McArthur, and Patricia Rodriguez-Gianolli. 2003. Answering clinical questions with role identification. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine*. Association for Computational Linguistics, pages 73–80.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2230–2235. http://aclweb.org/anthology/D16-1241.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1525–1534. https://doi.org/10.18653/v1/P16-1144.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 311–318. https://doi.org/10.3115/1073083.1073135.

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*. Association for Computational Linguistics, pages 2383–2392. https://doi.org/10.18653/v1/D16-1264.

Matthew Richardson, J.C. Christopher Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 193–203. http://aclweb.org/anthology/D13-1020.

Kirk Roberts, Matthew S Simpson, Ellen M Voorhees, and William R Hersh. 2015. Overview of the TREC 2015 Clinical Decision Support Track. In *TREC*.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17(5):507–513.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2017. CLAMP — a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association* https://doi.org/10.1093/jamia/ocx132.

Andreas Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *Proc. Int. Conf. Spoken Language Processing (ICSLP 2002)*.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 806–817. https://doi.org/10.18653/v1/P17-1075.

Richard FE Sutcliffe, Anselmo Peñas, Eduard H Hovy, Pamela Forner, Álvaro Rodrigo, Corina Forascu, Yassine Benajiba, and Petya Osenova. 2013. Overview of QA4MRE Main Task at CLEF 2013. In *CLEF (Working Notes)*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, pages 191–200. http://www.aclweb.org/anthology/W17-2623.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* 18(5):552–556.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 200–207.

Hai Wang, Takeshi Onishi, Kevin Gimpel, and David McAllester. 2017. Emergent predication structure in hidden state vectors of neural readers. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2604.

Yusuke Watanabe, Bhuwan Dhingra, and Ruslan Salakhutdinov. 2017. Question answering from unstructured text by retrieval and comprehension. *arXiv preprint arXiv:1703.08885* .

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017a. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. Association for Computational Linguistics, pages 94–106. http://www.aclweb.org/anthology/W17-4413.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2017b. Constructing datasets for multi-hop reading comprehension across documents. *arXiv preprint arXiv:1710.06481* .

A Training details and hyper-parameter optimization

We train the word embeddings using word2vec (Mikolov et al., 2013), and optimize the window size, the model type (CBOW, skip-gram), the dimensionality and the number of negative samples using random search. For the embedding baseline sim-entity, the evaluation was carried out 20 times on the validation part of our dataset, and we chose the parameter configuration that led to the highest-performing embedding model as measured by F1. We find that higher embedding dimensionality works better, that CBOW obtains somewhat better scores than Skipgram, and that medium-sized word windows work best. The best configuration: 'win_size': 5, 'min_freq': 200, 'model': 'cbow', 'dimension': 750, 'neg_samples': 5. The difference between the lowest and the highest scoring model was 3.4 F1. At prediction time (equation (1)) we set the window size to 3, which worked best on the validation set.

For inclusion in the neural readers, it would be impractical to use the high embedding dimensionality found in the hyper-parameter search from the previous paragraph, so we fix the input embedding dimensionality to 200, as done in Chen et al. (2016) to keep the training time practical. We optimize the remaining embedding hyper-parameters just like above. The best parameters were: 'win_size': 4, 'min_freq': 200, 'model': 'cbow', 'dimension': 200, 'neg_samples': 9.

For SA reader, we optimized the hidden state size and the dropout rate using 20 different random configurations. The best values were 70 and 0.57, respectively. We explore the same parameters for the GA reader, but add to the search space the feature that indicates the presence of a passage token in the query, which was found useful in the NoEnt set-up. The best hidden state number and dropout rate were 64 and 0.5, respectively. We used the default values for all the remaining hyper-parameters.

B List of skills with selected examples

In annotating the skills, we followed the categorization by Sugawara et al. (2017):

1. Object tracking: tracking or grasping multiple objects; it is a version of list/enumeration skill used in previous skill classifications
2. Mathematical reasoning: whenever a mathematical operation is involved in finding the answer
3. Coreference resolution: direct reference to an object, includes anaphoras. These include inferential processes based on background knowledge or context.
4. Logical reasoning: conditionals, quantifiers, negation, transitivity
5. Analogy: metaphors, metonymy
6. Causal relation: explicit expression such as "why", "the reason of"
7. Spatio-temporal relations
8. Ellipsis: recognizing implicit or omitted information
9. Bridging: inference through grammatical and lexical knowledge (synonymy, idioms etc). This link however is not automatic or stereotypical, as in the category of elaboration.
10. Elaboration: inference through commonsense reasoning. Note that unlike in the previous category, there is no direct way in which grammatical, lexical or ontological knowledge could help.
11. Meta-knowledge: knowing about the text genre and the main topic being discussed assists in comprehending. In our dataset, knowing the way the queries are constructed (Learning points) is sometimes beneficial.
12. Schematic clause relation: complex sentences that include coordination or subordination
13. Punctuation: understanding parentheses, dashes, quotations, colons etc.

In the following examples, we mark the medical entities in blue, and italicize the parts in the passage that are crucial for answering. Whenever we shorten a part of the passage, we use [...].

B.1 Bridging inference

passage

We report a case of a 72 - year - old Caucasian woman with pl-7 positive antisynthetase syndrome . Clinical presentation included interstitial lung disease , myositis , mechanic 's hands and dysphagia

. As lung injury was the main concern , treatment consisted of *prednisolone and cyclophosphamide* . Complete remission with reversal of pulmonary damage was achieved , as reported by CT scan , pulmonary function tests and functional status . [...]
query
Therefore , in severe cases an aggressive treatment , combining _____ and glucocorticoids as used in systemic vasculitis , is suggested .
answer
cyclophoshamide
explanation The reader needs to have the background knowledge that prednisolone is a glucocorticoid, then it becomes obvious that the answer is cyclophoshamide.

B.2 Object tracking

passage
[...] The patient was managed with supportive measures and the National Poisons Information Service was contacted . A toxicology consultant was involved in view of the unusual mode of administration . Although there was no precedent on how to treat a significant rectal overdose of amitriptyline , *it was advised that the patient be administered a phosphate enema and if failed to adequately remove the tablets then the patient should be given whole bowel irrigation* with 2 litre of Klean - Prep via a nasogastric tube . It was also advised that we admit the patient to a high dependency unit and manage him according to the usual protocol for a tricyclic overdose if complications arose . [...]
query
It seems reasonable to attempt careful removal of the drug from the rectum and if that fails to consider _____ and whole bowel irrigation .
answer
phosphate enemas
explanation The query mentions removal (A), then _____ (B) and whole bowel irrigation (C). In the passage, one needs to track those elements and choose the right one. This skill should be considered whenever the gap is part of an enumeration or is mentioned as a part of another entity.

B.3 Meta knowledge

query
bedaquiline , a new agent with bactericidal and sterilising activity against mycobacterium tuberculosis , is effective against _____ when given together with a background regimen , and is well tolerated and safe if there is awareness of drug interactions and precautions are taken to avoid potential qt prolongation .
answer
tuberculosis
explanation The right answer can be inferred from several parts in the passage (not shown), or even from the title or the query. The query, though, is nowhere in the document explicitly answered.