BioRead: A New Dataset for Biomedical Reading Comprehension

Dimitris Pappas[1,2], Ion Androutsopoulos[1], Haris Papageorgiou[2]
[1]Department of Informatics, Athens University of Economics and Business,
Patission 76, GR-104 34 Athens, Greece
[2]Institute for Language and Speech Processing, ATHENA Research Center,
Epidavrou & Artemidos 6, GR-151 25 Maroussi, Greece
pappasd@aueb.gr, ion@aueb.gr, xaris@ilsp.gr

Abstract

We present BioRead, a new publicly available cloze-style biomedical machine reading comprehension (MRC) dataset with approximately 16.4 million passage-question instances. BioRead was constructed in the same way as the widely used Children's Book Test and its extension BookTest, but using biomedical journal articles and employing MetaMap to identify UMLS concepts. BioRead is one of the largest MRC datasets, and currently the largest one in the biomedical domain. We also provide a subset of BioRead, BioReadLite, for research groups with fewer computational resources. We re-implemented and tested on BioReadLite two well-known MRC methods, AS Reader and AOA Reader, along with four baselines, as a first step towards a BioRead (and BioReadLite) leaderboard. AOA Reader is currently the best method on BioReadLite, with 51.19% test accuracy. Both AOA Reader and AS Reader outperform the baselines by a wide margin on the test subset of BioReadLite. Our re-implementations of the two MRC methods are also publicly available.

Keywords: BioRead, biomedical, dataset, corpus, evaluation, reading comprehension, question answering, deep learning.

1. Introduction

Machine Reading Comprehension (MRC) systems (Hermann et al., 2015) are given a passage (e.g., from a news article or book) and they are required to answer a question by considering the information of the passage. Manually constructing MRC datasets is very labour-intensive and leads to datasets that may not be large enough to train data-hungry deep learning methods (Goodfellow et al., 2016; Goldberg, 2017). For example, the BioASQ dataset (Tsatsaronis et al., 2015), which was constructed by biomedical experts, currently contains only two thousand questions approximately. The largest manually curated MRC dataset we are aware of, SQuAD (Rajpurkar et al., 2016), which was crowdsourced, comprises approximately 100k passage-question instances.

Larger datasets can be constructed automatically by considering only 'cloze-style' questions, which require filling in a missing word or phrase in a given sentence about the passage (e.g., "___ has been implicated in the pathogenesis of PD."). For example, CBTest (Hill et al., 2015) contains passages from children's books; each cloze-style question is a sentence that follows the corresponding passage in its book, with a randomly selected common noun, named entity, verb, or preposition of the sentence removed and turned into a slot to be filled in. CBTest contains approximately 687k passage-question instances. It was more recently expanded to BookTest (Bajgar et al., 2016), which comprises approximately 14 million passage-question instances, by applying the same methodology to a much larger collection of books. The CNN and Daily Mail datasets (Hermann et al., 2015) were produced in a similar manner. They comprise news articles and cloze-style questions constructed by removing words from sentences summarising the articles; they contain approx. 380k and 880k instances, respectively.

Apart from constituting a testbed for natural language understanding algorithms, MRC is also useful as a component of larger systems. We are interested in a setting where an Information Retrieval engine retrieves document passages that may be relevant to a question, and then MRC is used to identify exact answers (e.g., named entities) in the passages (Sultan et al., 2016; Chen et al., 2017). We focus on the biomedical domain, where this setting is included in the BioASQ challenges (Tsatsaronis et al., 2015). There is currently, however, no sufficiently large publicly available biomedical MRC dataset to train deep learning models. We, therefore, constructed and provide a new biomedical MRC dataset, called BioRead, with approx. 16.4 million cloze-style questions, each paired to a passage and candidate answers. BioRead was constructed in the same manner as CBTest and BookTest, using randomly selected biomedical articles from PubMed Central.[1] To the best of our knowledge, it is currently one of the largest MRC datasets, and the largest one in the biomedical domain. We also provide a subset of BioRead, called BioReadLite, with 900k instances, for groups with fewer computational resources.

We re-implemented (in PyTorch[2]), trained, and tested on BioReadLite two well-known MRC methods, AS Reader (Kadlec et al., 2016) and AOA Reader (Cui et al., 2017). We report their performance, along with the performance of four simpler baselines, as a first step towards a BioRead (and BioReadLite) leaderboard. We open-source the re-implementations to make it easier to replicate our experiments and build upon previous MRC methods.[3]

Automatically generated cloze-style MRC datasets are of lower quality compared to manually constructed ones. For example, Chen et al. (2016) reported that the CNN and Daily Mail datasets contain both questions that are too easy

[1]Consult `https://www.ncbi.nlm.nih.gov/pmc/`.

[2]See `http://pytorch.org/`.

[3]BioRead and the re-implementations will be made available at `http://nlp.cs.aueb.gr/software.html`. The original implementation of AS Reader (in Theano) is available at `https://github.com/rkadlec/asreader/`. The original implementation of AOA Reader does not appear to be online.

	BioRead				BioReadLite			
	Training	**Development**	**Test**	**Total**	**Training**	**Development**	**Test**	**Total**
Instances	∼15,1M	∼600,7k	∼652,9k	∼16.4M	800k	50k	50k	900k
Avg candidates	25.9	27.3	26.3	26.0	18.89	20.8	19.4	19.0
Max candidates	40	40	40	40	30	30	30	30
Min candidates	2	2	2	2	2	2	2	2
Avg context len.	456.9	464.5	455.9	457.1	317.2	320.8	298.9	316.4
Max context len.	999	999	999	999	400	400	400	400
Min context len.	26	56	48	26	30	30	30	30
Avg question len.	33.4	35.5	34.8	33.5	16.8	16.8	16.8	16.8
Max question len.	300	300	300	300	25	25	25	25
Min question len.	5	5	5	5	5	5	5	5

Table 1: Statistics of BioRead and BioReadLite. Lengths in tokens.

to answer using simple hand-crafted features (e.g., simple paraphrases of passage sentences) and questions that even people cannot answer. Nevertheless, BioRead (and BioReadLite) is the only sufficiently large biomedical MRC dataset to train deep learning methods on. In future work, we plan to investigate if systems trained (or pre-trained) on BioRead could also cope (possibly after further training) with real biomedical questions, like those of BioASQ.

2. The BioRead Dataset

To construct BioRead, we randomly selected approx. 90.6k from the approx. 3.4M articles (from approx. 7k biomedical journals) of the Open Access Subset of PubMed Central (PMC).[4] We then applied MetaMap (Aronson and Lang, 2010) to each one of the selected articles.[5] MetaMap recognises words or phrases referring to concepts of the Unified Medical Language System (UMLS).[6] As an example, the words and phrases shown in red or green in the 'context' of the left column of Table 2 were recognised as UMLS concepts. MetaMap also provides the 'preferred name' of each concept. For example, 'carcinoma of the lung', 'lung cancer', and 'malignant tumor of the lung' all refer to the same concept; the preferred name is the first one.

To reduce the size of the vocabulary and avoid confusing MRC methods by synonyms, we replaced each concept that MetaMap recognized by its preferred name. Borrowing the notation of the CNN and Daily Mail datasets (Hermann et al., 2015), each preferred name (possibly multi-token) was then mapped to a pseudo-token of the form @entityID (Fig. 2, right), where ID is an integer identifier that is (a) unique within the particular passage-question instance (i.e., the same ID will generally denote a different concept in another instance), or (b) unique in the entire dataset (the same ID will always denote the same concept). Hermann et al. (2015) ensure that expressions referring to the same entity get the same ID within the same passage-question instance only, which corresponds to option (a). This does not allow systems to learn information about an entity from multiple passages of the training set. By contrast, in option (b), where each entity (concept) has the same ID in the entire dataset, an MRC method may, at least in principle, learn properties of an entity from the passages of multiple training instances. We adopt Hermann et al.'s option (a) in most of our experiments, but we also train the best method with option (b), which improves performance.

As in CBTest and BookTest (Hill et al., 2015; Bajgar et al., 2016), having replaced the recognized entities (concepts) by @entityID pseudo-tokens, we applied a sliding window of 21 sentences to the texts of the approx. 90.6k articles.[7] For each position of the window, we examined each @entityID of the 21st sentence. If an @entityID of the 21st sentence was also present (with the same ID) anywhere in the first 20 sentences, that @entityID was replaced by a @placeholder pseudo-token in the 21st sentence, indicating a slot to be filled in by one of the @entityID tokens of the first 20 sentences; the 21st sentence became a cloze-style question (Fig. 2, right), the first 20 sentences became the 'context' (passage) of the question, the @entityID tokens of the context became the candidate answers, and the particular @entityID that was turned into @placeholder became the correct answer. If multiple @entityID tokens of the 21st sentence were present in the first 20 sentences, multiple context-question-candidates-answer tuples were obtained; hence, strictly speaking each *instance* of BioRead is a context-question-candidates-answer tuple, not just a context and question pair. If no @entityID tokens of the 21st sentence were present in the first 20 sentences, no instance was produced from that position of the window.

To lower the computational resources required to process the dataset, we also set a maximum context length of 999 tokens, a maximum question length of 300 tokens, and a maximum of 40 candidate answers per instance. We discarded instances exceeding these thresholds, obtaining approx. 16.4M instances from the approx. 90.6k articles. BioRead contains these instances, split into training, development, and test sets (Table 1). We also provide a subset of BioRead, called BioReadLite, for researchers with fewer computational resources. BioReadLite was created by setting the maximum context length to 400 tokens, the maximum question length to 25 tokens, and the maximum number of candidate answers to 30. The vocabulary of each dataset includes all the words (and pseudo-tokens) that occurred at least 5 times in the corresponding training subset,

[4] Consult `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`. The articles were available in plain text and HTML format; we used the former.

[5] See `https://metamap.nlm.nih.gov/`.

[6] See `https://www.nlm.nih.gov/research/umls`.

[7] We used NLTK's sentence splitter (`http://www.nltk.org/`) and a basic white-space tokeniser.

Context: *salsolinol* (100mg/kg i.p.) or *l-dopa* (100mg/kg i.p.) was acutely administered (100mg/kg i.p.). in the combined treatment group, *l-dopa* (100mg/kg i.p.) was administered once 15min after *salsolinol* administration. the *rats* were decapitated 2h after *injection*. the *concentration* of *dopamine* and its metabolites were measured using *hplc* . the results are expressed as the means sem (n=710 animals per group). the data were analyzed via two-way anova followed by duncans test. statistical significance: [...] *l-dopa* (f[1,27]=26.9, p<0.01) on the level of 3-mt (table1). however, neither *treatment* with *salsolinol* (f[1,27]=0.09, n.s.) nor the interaction between *salsolinol* and *l-dopa* (f[1,27]=0.03, n.s.) was significant (table1).	**Context:** *@entity10* (100mg/kg i.p.) or *@entity5* (100mg/kg i.p.) was acutely administered (100mg/kg i.p.). in the combined treatment group, *@entity5* (100mg/kg i.p.) was administered once 15min after *@entity10* administration. the *@entity6* were decapitated 2h after *@entity0* the *@entity3* of *@entity7* and its metabolites were measured using *@entity9* the results are expressed as the means sem (n=710 animals per group). the data were analyzed via two-way anova followed by duncans test. statistical significance: [...] *@entity5* (f[1,27]=26.9, p<0.01) on the level of 3-mt (table1). however, neither *@entity2* with *@entity10* (f[1,27]=0.09, n.s.) nor the interaction between *@entity10* and *@entity5* (f[1,27]=0.03, n.s.) was significant (table1).
Question: the duncans post hoc test showed that *l-dopa* induced an increase in the *concentration* of 3-mt (by approximately 300%, p<0.01) but that *salsolinol* did not influence this effect of *l-dopa* (table1).	**Question:** the duncans post hoc test showed that *@placeholder* induced an increase in the *@entity3* of 3-mt (by approximately 300%, p<0.01) but that *@entity10* did not influence this effect of *@entity5* (table1).
Candidates: injection, control group, treatment, concentration, substantia nigra, l-dopa, rats, dopamine, dopac, hplc, salsolinol, analysis	**Candidates:** @entity0, @entity1, @entity2, @entity3, @entity4, @entity5, @entity6, @entity7, @entity8, @entity9, @entity10, @entity11
Answer: l-dopa	**Answer:** @entity5

Table 2: An example instance of BioRead, before (left) and after (right) replacing recognized UMLS concepts by pseudo-tokens. Red words and phrases are wrong candidate answers. The correct answer is shown in green and underlined.

after replacing all digits with 'D' (e.g., 'type-3' becomes 'type-D'). The resulting vocabulary sizes of BioRead and BioReadLite are approx. 3.9M and 597k, respectively. Out-of-vocabulary words have been replaced by 'UNK'.

By replacing (possibly multi-token) concept names with @entityID tokens, we allow MRC methods that can only select a single token from the passage (the two methods we re-implemented belong in this category) to cope with cases where the correct answer is actually multi-token. Furthermore, by replacing concept names with @entityID, we do not let MRC systems look up the concepts in external resources (e.g., biomedical ontologies), forcing them to base their responses on the passages of the dataset. The use of MetaMap, however, also adds noise, since MetaMap is not entirely accurate. For example, in Fig. 2 (left), it failed to recognise 'metabolites' as a biomedical concept.[8] Similar noise was introduced in CBTest and BookTest by the named entity recognisers that were used during their construction.

3. Re-implemented Methods and Baselines

We re-implemented and experimented with AS Reader (Kadlec et al., 2016), because it is one of the simplest and most well-known deep learning MRC methods. It has also been shown (Bajgar et al., 2016) that increasing the size of the training set of AS Reader (using BookTest instead of CBTest) leads to much larger performance gains than training more complex MRC methods, like AOA Reader (Cui et al., 2017) and EpiReader (Trischler et al., 2016), on the original training set (CBTest). We also reimplemented and experimented with AOA Reader (Cui et al., 2017), an extension of AS Reader that uses a more complex attention mechanism, because it is one of the best performing methods on CBTest (Bajgar et al., 2016). We make both re-implementations publicly available, as already noted.

AS Reader (Kadlec et al., 2016) uses a bidirectional recurrent neural network (biRNN) (Schuster and Paliwal, 1997; Seo et al., 2016) with GRU units (Cho et al., 2014) to process the passage (context) and another one to process the question. The states of the first biRNN (the concatenated states of the two directions, for each token position) are used as context-sensitive embeddings of the passage tokens, whereas the last states of the second biRNN (the concatenated last states of the two directions) represent the question. The dot product between the question representation and the context-sensitive embedding of each passage token is then computed, and a softmax is applied to the dot products to turn them into attention scores from 0 to 1. The candidate answers can only be single tokens of the passage. If a candidate answer occurs multiple times in the passage, its attention scores are summed. Finally, the candidate answer with the largest (summed) attention score is selected.

AOA Reader (Cui et al., 2017) uses a biRNN to create context-sensitive embeddings for each passage token, as in AS Reader. Another biRNN processes the question, but instead of keeping only the (concatenated) last states of the two directions as the question representation, all the states

[8] We configured MetaMap for high precision, by setting its minimum score of recognised concepts to 10.

of the question biRNN (the concatenated states from both directions, for each question token position) are kept as context-sensitive embeddings of the question tokens. The dot product between each context-sensitive embedding of the passage and each context-sensitive embedding of the question is then computed, leading to a matrix $M = C \times Q$ of dot products, where C and Q are the lengths of the passage (context) and question, respectively, in tokens. Intuitively, each element $m_{i,j}$ of M shows how relevant token i of the passage is to token j of the question. The i-th row of M contains Q scores, showing how relevant each token of the question is, from the viewpoint of the i-th token of the passage. The rows of M are averaged (after applying a softmax to each row first) to obtain a single row-vector q with Q scores that shows how relevant each token of the question is with respect to *all* the tokens of the passage.
Similarly, the j-th column of M contains C scores showing how relevant each token of the passage is, from the viewpoint of the j-th token of the question. The matrix-vector multiplication $M'q^T$, where M' is the original M with a softmax applied to each column, produces C scores that show how important each passage token is from the viewpoint of the entire question, as captured by q. A softmax is applied to the C scores, to turn them into attention scores from 0 to 1. As in AS Reader, the candidate answers can only be single tokens of the passage. If a candidate answer occurs multiple times in the passage, its attention scores are summed. Finally, the candidate answer with the largest (summed) attention score is selected.

Baselines: The first baseline, called BASE1, returns the candidate answer (@entityID) that occurs most frequently in the context (passage), on the grounds that this candidate answer is more likely to have also occurred in the question (a sentence that follows the passage) and, hence, more likely to have been converted to @placeholder. The second and third baselines, BASE2 and BASE3, return the candidate answer that occurs first or last in the context, respectively. The last candidate answer is arguably more likely to be repeated in the question and, hence, more likely to have been converted to @placeholder, whereas the first candidate is the least likely to be repeated in that sense. We also suspected that the biRNN encoder of the passage of AS Reader and AOA Reader would tend to 'remember' more the last (in the forward RNN) and the first (in the backwards RNN) tokens (and candidate answers) of the passage, in an extreme case behaving like BASE2 and BASE3.
In the fourth baseline, BASE4, we first extract all the token n-grams ($n = 2$) of the question that contain the @placeholder.[9] For each candidate answer (@entityID), we then replace the @placeholder in all the extracted n-grams by the particular candidate answer, and count the total number of occurrences of the resulting n-grams in the context. The candidate answer with the largest total number of n-gram occurrences is returned as the answer.

Human performance: To get a rough estimate of how easily humans can answer the questions of BioRead, we randomly selected 30 instances from BioRead's test subset and gave them to three human annotators (the first two authors and a colleague), who had no biomedical background. The annotators were shown the context and question of each instance (as in Fig. 2, right) in a user interface that displayed @entityIDs as hyperlinks, and they were asked to select (click on) the correct candidate answer (@entityID). When the annotators felt they were clueless (or very uncertain) about the correct answer, they could indicate this by clicking on a button, but they were instructed to select an answer when they felt it was probably the correct one, even if they were not entirely sure. The mean accuracy of the three annotators was 68.01% (77.27%, 65.22%, 61.54% per annotator), counting only instances they answered (78.89% on average, 73.33%, 76.67%, 86.67% per annotator). The mean pairwise inter-annotator agreement, measured as Cohen's Kappa (Cohen, 1960), was 68.57, considering only questions answered by both annotators in each pair. If not answering a question is treated as an additional candidate answer, the mean pairwise Kappa becomes 50.32.

4. Experimental Results

Table 3 summarises our experimental results on BioRead-Lite; we did not have the computational resources to experiment with the full BioRead dataset, but we hope that others may be able to do so.[10] With the exception of the last row of Table 3, in all other cases we used option (a) of Section 2, i.e., the identifier of each @entityID was unique only within the particular instance. For AS Reader and AOA Reader, we used the same hyper-parameter values as in the work of Kadlec et al. (2016) and Cui et al. (2017), respectively. Hence, a direct possible improvement would be to fine-tune the hyper-parameters for BioRead (or BioRead-Lite), which requires, however, substantial computational resources. We stopped training the two methods when their development loss had converged, i.e., after 5 epochs for AS Reader, 15 epochs for AOA Reader when using option (a), and 20 epochs for AOA Reader when using option (b); recall that in option (b) the identifier of each @entityID is unique in the entire dataset. A single training epoch (including computing the development loss) takes 17, 21, and 22 hours, respectively (Table 3). Performance is measured in terms of accuracy, i.e., number of correctly answered development or test instances, divided by the total number of development or test instances.
Table 3 shows that AOA Reader is clearly more accurate than AS Reader, at the expense of training speed, reaching 50.44% and 49.94% development and test accuracies with option (a), compared to 37.90% and 42.01% for AS Reader, respectively. These results confirm that the more elaborate attention mechanism of AOA Reader is important, as also reported in previous work (Cui et al., 2017; Bajgar et al., 2016; Munkhdalai and Yu, 2016). Despite its simplicity, BASE1 (most frequent candidate answer in the passage) is a reasonably strong baseline, reaching 26.86% development and 28.87% test accuracy, but AS Reader and AOA Reader outperform it by a wide margin. BASE2 and BASE3 are much weaker, suggesting that AS Reader and

[9] We experimented with $2 \leq n \leq 6$, and selected $n = 2$, which led to the best results on the development set of BioReadLite.

[10] We used a PC running Ubuntu, with 64 GB RAM, a 16 core CPU, and a GeForce GTX TITAN X GPU with 12GB memory.

Method	Dev. Accuracy	Test Accuracy	Training Epochs
BASE1 (a)	26.86	28.87	n/a
BASE2 (a)	8.14	9.38	n/a
BASE3 (a)	16.48	17.28	n/a
BASE4 (a)	40.10	37.20	n/a
AS Reader (a)	37.90	42.01	5 × 17 h
AOA Reader (a)	50.44	49.94	15 × 21 h
AOA Reader (b)	**52.41**	**51.19**	**20 × 22 h**

Table 3: BioReadLite results (%), and number of epochs (and time) required for the development loss to converge, when each entity ID is unique (a) in the particular instance only, or (b) in the entire dataset.

AOA Reader do not just remember the first or last candidate answers of the passage. The best baseline is BASE4 (n-grams). It scored 40.10% development and 37.20% test accuracy, surpassing AS Reader on the development subset, and challenging AS Reader on the test subset. Nevertheless, AOA Reader outperformed BASE4 by a wide margin (Table 3). The performance of AOA Reader improved further (from 50.44% to 52.41% development accuracy, from 49.94% to 51.19% test accuracy), at the expense of additional training time, when option (b) was used, i.e., when each entity ID was unique in the entire dataset, suggesting that AOA Reader was able to learn properties of at least some entities (concepts) from multiple training passages.

We also trained the best method, AOA Reader with option (b), on smaller subsets of BioReadLite to study the effect of the size of the training set. We always used 20 epochs in this experiment, the number of epochs it took for the development loss to converge when using the entire training set of BioReadLite (Table 3, last row). Table 4 shows that increasing the size of the training set leads to improved development accuracy. We see a similar trend in test accuracy (from 49.22% to 51.19%) when going from 50% to 100% of the training set, but surprisingly the best test accuracy (51.51%) was obtained when using only 25% of the training set. The latter may be the result of a random fluctuation (e.g., the optimizer may have managed to find a better local minimum of the loss function in that case). It would be better to repeat each experiment multiple times, with different random parameter initializations, and report mean results (and standard deviations), but we did not have the required resources. Overall, however, it seems worth experimenting with the entire BioRead dataset, instead of BioReadLite, to see if its larger training subset would lead to significant improvements in accuracy. We also note that the average accuracy of the human annotators was 68.01% (Section 3). This score was computed only on a sample of 30 test questions, and it does not consider questions the annotators left unanswered, but it is an indication that there is headroom for improvements in the performance of MRC methods.

Training Subset	Dev. Accuracy	Test Accuracy	Training Epochs
25%	47.06	**51.52**	20 × 6 h
50%	50.25	49.22	20 × 11 h
100%	**52.41**	51.19	20 × 22 h

Table 4: BioReadLite results (%) of AOA Reader, with option (b), using the entire or only subsets of the training set.

5. Conclusions and Future Work

We constructed and make publicly available a new cloze-style biomedical MRC dataset, BioRead, with approx. 16.4 million instances, currently one of the largest MRC datasets and the only one of its kind in the biomedical domain. We also provide a subset of BioRead, BioReadLite, with 900k instances, for groups with fewer resources. Both datasets were constructed in the same way as CBTest and BookTest, but using biomedical journal articles and employing MetaMap to identify biomedical entities (concepts) and replace them by their preferred UMLS names. We also re-implemented and tested on BioReadLite two well-known MRC methods, AS Reader (Kadlec et al., 2016) and AOA Reader (Cui et al., 2017), along with four baselines, as a first step towards a BioRead (and BioReadLite) leaderboard. Our re-implementations are also publicly available.

BioRead and BioReadLite are available in two forms, where each identified entity is replaced by an identifier that is unique (a) only in the particular passage-question instance, or (b) in the entire training corpus. AOA Reader is currently the best method on BioReadLite, and its performance improves (reaching 52.41% development and 51.19% test accuracy) when option (b) is used, suggesting that it manages to learn properties of at least some entities from multiple training passages. The best baseline, which uses n-grams, surpasses the second best method, AOA Reader, on the development set of BioReadLite and performs reasonably well on the test set. Nevertheless, AOA Reader outperforms it by a wide margin.

Future work could use BioRead (and BioReadLite) to test other existing MRC methods in the biomedical domain or develop new MRC methods. It would also be interesting to examine if methods trained (or pre-trained) on BioRead could also cope with real-world biomedical questions, like those of BioASQ, possibly after further training.

6. Acknowledgments

We acknowledge support of this work by the project "Computational Science and Technologies: Data, Content and Interaction" (MIS 5002437) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund). We are also grateful to Ryan McDonald, for many useful discussions and his participation in the human performance experiment. Special thanks to Yannis Almirantis and Anastasios Nentidis, who participated in a similar preliminary experiment with biomedical experts.

7. Bibliographical References

Aronson, A. R. and Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17:229–236.

Bajgar, O., Kadlec, R., and Kleindienst, J. (2016). Embracing Data Abundance: BookTest Dataset for Reading Comprehension. *CoRR*, abs/1610.00956.

Chen, D., Bolton, J., and Manning, C. D. (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*, pages 2358–2367, Berlin, Germany.

Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017). Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1870–1879, Vancouver, Canada.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., and Hu, G. (2017). Attention-over-Attention Neural Networks for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 593–602, Vancouver, Canada.

Goldberg, Y. (2017). *Neural Network Methods in Natural Language Processing*. Morgan and Claypool Publishers.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, pages 1693–1701, Montreal, Quebec, Canada.

Hill, F., Bordes, A., Chopra, S., and Weston, J. (2015). The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *CoRR*, abs/1511.02301.

Kadlec, R., Schmid, M., Bajgar, O., and Kleindienst, J. (2016). Text Understanding with the Attention Sum Reader Network. *CoRR*, abs/1603.01547.

Munkhdalai, T. and Yu, H. (2016). Reasoning with Memory Augmented Neural Networks for Language Comprehension. *CoRR*, abs/1610.06454.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas.

Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Seo, M. J., Kembhavi, A., Farhadi, A., and Hajishirzi, H. (2016). Bidirectional Attention Flow for Machine Comprehension. *CoRR*, abs/1611.01603.

Sultan, M. A., Castelli, V., and Florian, R. (2016). A Joint Model for Answer Sentence Ranking and Answer Extraction. *Transactions of the Association of Computational Linguistics*, 4:113–125.

Trischler, A., Ye, Z., Yuan, X., Bachman, P., Sordoni, A., and Suleman, K. (2016). Natural language comprehension with the epireader. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Austin, Texas.

Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artieres, T., Ngonga, A., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., and Paliouras, G. (2015). An overview of the BioASQ Large-Scale Biomedical Semantic Indexing and Question Answering Competition. *BMC Bioinformatics*, 16(138).