Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-41

A Question Answering System on Regulatory Documents

Diego COLLARANA [a,c,1], Timm HEUSS [b], Jens LEHMANN [a,c], Ioanna LYTRA [a,c], Gaurav MAHESHWARI [a,c], Rostislav NEDELCHEV [a,c] Thorsten SCHMIDT [b] and Priyansh TRIVEDI [a,c]

[a] *Enterprise Information Systems, Fraunhofer IAIS, Germany*
[b] *PricewaterhouseCoopers GmbH, Germany*
[c] *Smart Data Analytics Group, University of Bonn, Germany*

Abstract. In this work, we outline an approach for question answering over regulatory documents. In contrast to traditional means to access information in the domain, the proposed system attempts to deliver an accurate and precise answer to user queries. This is accomplished by a two-step approach which first selects relevant paragraphs given a question; and then compares the selected paragraph with user query to predict a span in the paragraph as the answer. We employ neural network based solutions for each step, and compare them with existing, and alternate baselines. We perform our evaluations with a gold-standard benchmark comprising over 600 questions on the MaRisk regulatory document. In our experiments, we observe that our proposed system outperforms other baselines.

Keywords. Question Answering; Reading Comprehension; Regulatory Domain

1. Introduction

Users of legal or regulatory documents usually access information through a keyword based search over a collection of related documents. Information Retrieval (IR) techniques can improve the search results by making use of synonyms, ontologies, and word embeddings, which allows an intelligent lookup rather than a simple keyword-based search. Still this approach requires the commitment of the legal expert in examining the retrieved documents and/or excerpts in order to find the right information providing answers to her queries. There is, hence, an increasing interest in developing systems able to provide accurate and precise answers to a user query, enabling a more intuitive, fast, and accurate access to information. One way for achieving this is the use of a question answering (QA) systems on such documents which can provide direct answers to users' questions by highlighting text spans in the target documents.

Question Answering is an important topic of research in the natural language processing field and and has seen major advances in the last years. Research in this area dates back to the 1960s with recent popular milestones being the development of IBM

[1]Corresponding Author: Enterprise Information Systems, Fraunhofer IAIS, Schloss Birlinghoven 53757 Sankt Augustin Germany; E-mail: diego.collarana.vargas@iais.fraunhofer.de.

Watson [4] and various conversational agents such as Alexa, Google Assistant, Siri, etc. Recent reading comprehension datasets like [12,10] further accelerated the growth of the field, by providing largescale general-domain datasets, enabling the use of complex neural network based approaches [16,7]. However, porting these techniques to domain-specific applications is non-trivial, and only few solutions [11,2] have been proposed for QA over legal documents. Amongst them, to the best of our knowledge, no comparable approach or system exists for the regulatory domain. This is mainly due to the fact that the specific domain – the regulatory domain – poses several challenges which need to be overcome in order to design and implement an effective QA system: (1) *Existence of long documents*: documents in the regulatory domain are usually long, consisting of tens or hundreds of pages, increasing the search space for potential answers; (2) *Structure of documents*: generally, regulatory documents have a rather complex structure including sections, subsections, paragraphs; (3) *Use of domain-specific language*: regulatory documents use a domain-specific vocabulary and a complex language which is difficult to be interpreted by machines. Further, the meaning of the words and the contexts often differs from the general domain or even across different regulatory documents; (4) *Lack of training data in the domain*: although several benchmarks for the general domain are available (e.g., SQuAD [12], MS MARCO [10], etc.), domain-specific datasets are very rare. Therefore, the key to a robust QA system over regulatory texts is a system that can comprehend both coarse and fine-grained information (across long documents) and adapt to the domain-specific vocabulary and complex language used in such documents.

In this paper, we propose a QA system over the regulatory domain which retrieves the paragraph and the precise span of the document as the answers to a user query. It consists of two major modules, namely *paragraph selection module*, which finds the relevant paragraph in the document given a question, and *answer selection module* which given a paragraph points to the exact span in the document. We show empirically, that our proposed approach can be effective for retrieving the answer span and outperforms traditional IR techniques. Further, we employ transfer learning techniques to increase the model performance, given the lack of training data in our specific domain. The proposed approach has been evaluated with question-answer pairs for the English translation of the MaRisk regulations document[2]. We tested different models and configurations of the proposed approach which allowed us to compare alternative QA pipelines.

The remainder of the paper is structured as follows. We introduce the difference between a traditional IR approach and a QA system based on reading comprehension through a motivating example in Section 2. The proposed approach is described in detail in Section 3 and the evaluation results are discussed in Section 4. Section 5 discusses the related work and, finally, Section 6 summarizes our key conclusions.

2. Motivating Example

In this section, we introduce a example of a question and answers based on the MaRisk regulatory document which defines Minimum Requirements for Risk Management for banks, insurances, and companies financially trading in Germany. The 62-page long MaRisk document in English consists of 2 main parts (General Part and Special Part),

[2] https://www.bafin.de/SharedDocs/Veroeffentlichungen/EN/Meldung/2014/meldung_140815_marisk_uebersetzung_en.html, last accessed on 2018-09-18

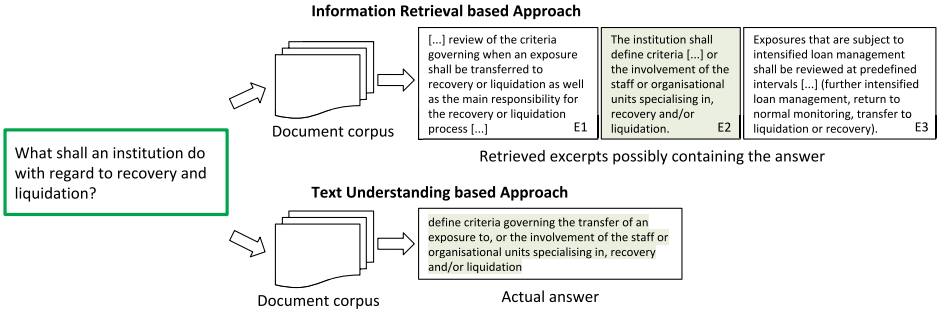

Figure 1. Comparison of IR approach for retrieving answers to natural language questions to the QA approach based on reading comprehension. While with the first, we retrieve text passages which appear to be "similar" to the question, in the second approach the text excerpt is semantically mapped to the question.

64 sections and subsections having several paragraphs among which many are supported by annotations. It has approximately 24,000 words with many of them in German. A few representative questions are the following: "Where shall it be ensured that a trader can enter trades only under his/her own trader ID?", "Why are reverse stress tests performed?", and "What shall an institution do with regard to recovery and liquidation?". The spectrum of questions which can be asked is very broad (what-, how-, why-questions, etc.) and the answers may vary from a few words to a sentence, several sentences, or a paragraph. Figure 1 visualizes two alternative approaches for retrieving an answer from a document corpus given a question expressed in natural language: a traditional IR approach and a QA approach based on reading comprehension which is being investigated in the current work. While the first will return possible excerpts containing answer(s) to the user's question, the latter will find the actual answer to the initial question. For instance, given the exemplary question *"What shall an institution do with regard to recovery and liquidation?"*, based on the similarity of the question to several paragraphs in the MaRisk document, several excerpts with high confidence are being retrieved, with some of them (E1 and E3) not being related to the intent of the question. However, a reading comprehension based method is expected to interpret the question and find the actual answer, based on a deeper, albeit implicit understanding of the document.

3. Proposed Approach

To answer questions on long regulatory text documents we propose a two-step approach: 1. *Paragraph selection (PS)*: The purpose of this step is to find the most relevant paragraphs of a long document so that (costly) state-of-the-art reading comprehension techniques can be efficiently applied on the extracted paragraph(s). 2. *Answer Selection (AS)*: Given a paragraph retrieved by Step 1, the aim of this step is to find the exact word span which provides the answer to the question. The following subsections describe the details of paragraph and answer selection.

3.1. Paragraph Selection

The first phase of the pipeline involves selecting the relevant paragraph given a question. To do so, we employ an approach described in [17]. The proposed model, StarSpace

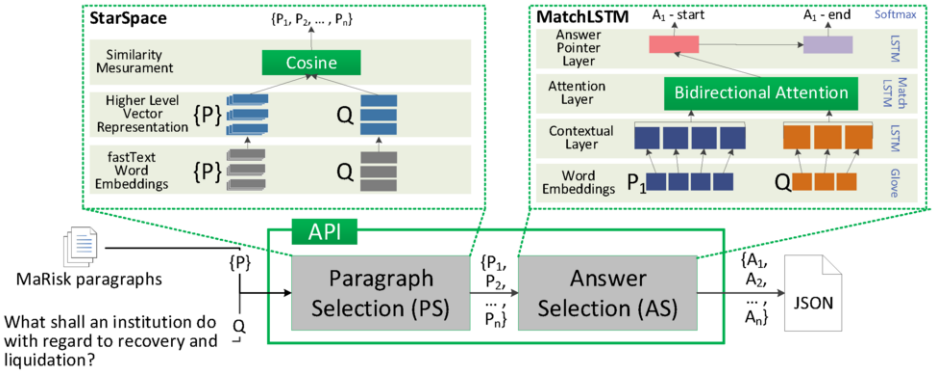

Figure 2. A two-step question answering approach for regulatory documents comprising: (1) Paragraph Selection and (2) Answer Selection.

(i) uses negative sampling to minimize margins between related entities, and maximizes margins between queries and irrelevant text, and (ii) is able to compare unrelated entities. This is accomplished by using a hierarchical structure of features to describe its entities, allowing to "embed all the things".

In our implementation, words in both, queries, and documents are numerically represented by embeddings provided by fastText [8]. To generate positive pairs, we broke each of the documents into words. Then we generated n-grams of different sizes (n=3,4,5,6) that we used as a "pseudo-query" for each document. For a wider understanding of difficulty of the task, we also employ a Lucene index in parallel to select the paragraphs. The performance of both approaches can be found in Section 4.

3.2. Answer Selection

In order to find the exact span of the answer in the paragraph, retrieved by the Paragraph Selection module, we use the architecture proposed by [16]. The architecture consists of two major layers, a Match-LSTM layer and an Answer-Pointer layer. The Match-LSTM layer employs a word-by-word cross-attention to form a new weighted version of the paragraph representation based on the input question. This new representation indicates the degree of matching of each word in the paragraph to that of the question. This new weighted representation is then combined with the original question vector to form the final representation of the question which acts as an input to the Answer-Pointer layer. The Answer-Pointer layer employs a self-attention based mechanism to point to the exact span in the paragraph. The primary reason of using this architecture is that it points back to the span, rather than generating the answer on its own. This makes the architecture more robust to domain specific words, as it does not need to generate them from scratch.

Due to its complexity, and the inherent difficulty of the task, the model needs a large amount of data to train on. However, in our domain, supervised data is sparse and difficult to generate. To offset this general lack of data, we first train our model on SQuAD [12], and then fine-tune it over the domain dataset. Fine tuning is necessary as the underlying characteristics of one dataset is usually different from another. For example, while SQuAD has on average 11.31 words in the question and 2.47 words in the answer, our benchmark consists of questions including one more word on average and significantly longer answers (more than 10 words).

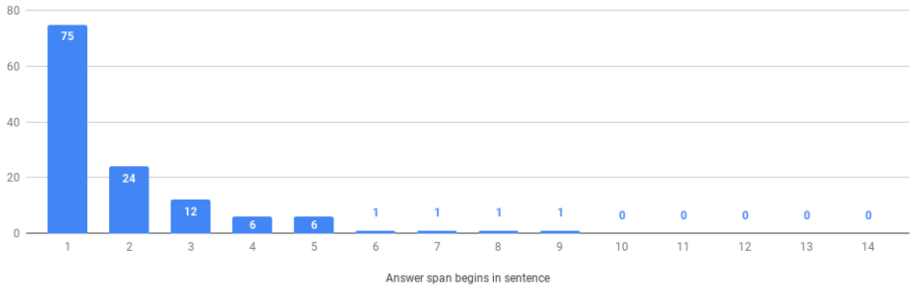

Figure 3. Distribution of the *start* of answers in MaRisk across paragraph sentences.

3.3. Question Generation

Contemporary machine learning models require copious amounts of data to achieve optimal results. And even when trained properly, their performance is highly representative of the inherent characters of the dataset. In our use case, we found out that the dataset is too small and disproportionately focuses on the first sentence of paragraphs. Also, models trained on this dataset perform poorly and do not generalize well.

To alleviate these issues, apart from using transfer learning, as described above, we also synthetically expanded the training data. We made use of a neural network question generation model proposed by [3]. The model is first trained on SQuAD [12]. We then apply the proposed method over the regulatory documents to generate a question for each of its sentences. In this manner, we overcome a major challenge hindering the use of statistical models for any subtask in the pipeline.

4. Evaluation Results

Dataset The ground truth consists of 631 question-answer pairs based on the MaRisk document. As mentioned above, a big majority of the questions have their answers in the beginning of the paragraph (first sentence), as visualized in Figure 3. On average, questions and answers have 12 and 13 words respectively. This is unlike SQuAD [12], whose answer spans are only three words long, on average. This distinction is noteworthy, as it makes transferring trained models from SQuAD to our dataset more challenging.

Metrics We report four metrics to evaluate the proposed approach and compare its effectiveness to alternative approaches: *Precision (P)*: ratio of the correct part of the predicted answer to the length of the predicted answer. *Recall (R)*: ratio of the the correct part of the predicted answer to the length of the true answer. *Partial Overlap (F1)*: Harmonic mean of P and R. *Exact Match (EM)*: This metric measures the percentage of predictions that exactly match the ground truth.

4.1. Answer Selection Evaluation

The answer selection module is trained in three different configurations, namely (i) on SQuAD without fine-tuning on the MaRisk data, (ii) on the MaRisk data (iii) pre-trained

	F1	EM
Trained on SQuAD	0.36	0.06
Trained on In-Domain data	0.20	0.10
Trained on SQuAD + In-Domain data	**0.59**	**0.34**

Table 1. Comparison of performance of the Answer Selection module trained in three different configurations: (i) only on the SQuAD dataset, (ii) on in-domain data, and (iii) on SQuAD and fine-tuned on in-domain data. In the third case, we are able to achieve significantly better results.

on SQuAD and then fine-tuned on the MaRisk data. The performance of the answer selection module trained in these configurations is summarized in Table 1.

We observe that training the model on SQuAD without any fine-tuning on the target dataset leads to very low performance. We attribute this to the difference in the inherent dataset characteristics. Moreover, while SQuAD consists of general domain text, the target task has a specialized vocabulary, which might impart specific, and sometimes completely different meaning to its constituent words. Also, the model performs almost equally worse when just trained on the domain dataset. This is as expected, as due to the limited size of the dataset the model fails to generalize on unseen data. However, when trained on SQuAD and afterwards fine-tuned in the regulatory domain, the results improve significantly. This suggests that transfer learning can, in general, improve the performance of our models, thereby helping to overcome the problem of the low number of training data in the specific domain.

4.2. Overall System Evaluation

The results of the QA system based on the two-step approach (Paragraph Selection and Answer Selection modules) are presented in Table 2. In particular, we compare four different pipelines composed by the combination of two variants for paragraph selection and two for answer selection. For paragraph selection, we compare StarSpace based model to a simple Lucene index (hereafter referred to as IR approach). For answer selection, we compare the MatchLSTM model, to a sentence selection based QA model, as proposed in [1].

In our experiments, we conclude that the IR based paragraph retrieval, and MatchLSTM based answer retrieval pipeline delivered the best results, across all the reported metrics. However, using StarSpace for paragraph retrieval instead of the IR approach also delivers comparable results. In addition, we observe that the sentence selection based approach for answer span selection performs worse. This is as expected since the answer spans in MaRisk do not follow sentence boundaries.

Instead of selecting the best paragraph in the first step, we also experiment with predicting more than one probable candidate paragraphs. We observe that the pipeline StarSpace (PS) + MatchLSTM (AS) performs the best amongst the baselines when k=3. However, when k=5, IR (PS) + MatchLSTM (AS) outperforms it. In general, the two paragraph selection algorithms can be used interchangeably since they deliver very similar results. In fact, the IR based approach achieves F1-measure and exact match only by 0.045 and 0.065 respectively higher compared to StarSpace when k=1.

Table 2. Evaluation of the proposed models. We report the Precision (P), Recall (R), F1-measure (F1) and Exact Match (EM). The metrics report accuracy over the final task of selecting the correct answer span.

QA Pipeline	P	R	F1	EM
IR (PS) + Choi et al. (SS)	0.182	0.479	0.263	0.000
IR (PS) + MatchLSTM (AS)	**0.667**	**0.574**	**0.617**	**0.343**
StarSpace (PS) + Choi et al. (SS)	0.182	0.446	0.259	0.000
StarSpace (PS) + MatchLSTM (AS)	0.618	0.532	0.572	0.278

A. Performance of baselines and the proposed model. Best performances are in **bold font**.

QA Pipeline	Top-k	P	R	F1	EM
IR (PS) + MatchLSTM (AS)	Top-1	0.667	0.574	0.617	0.343
	Top-3	0.702	0.617	0.657	0.343
	Top-5	**0.734**	**0.646**	**0.688**	**0.352**
StarSpace (PS) + MatchLSTM (AS)	Top-1	0.618	0.532	0.572	0.278
	Top-3	0.713	0.616	0.661	0.333
	Top-5	**0.74**	**0.641**	**0.687**	**0.343**

B. Two best performing models when considering top-1, top-3 and top-5 paragraphs for the paragraph selection step. Best performances are in **bold font**.

4.3. Discussion

The evaluation results suggest the following conclusions:

- Apart from the size of the dataset, its balance (in our case, position of the answer span) is pivotal for robust generalizations. In our approach, as mentioned in Section 3.3, we attempt to balance the dataset by synthetically generating questions from all the sentences in the document. We find out that synthetic data increases the model performance by 11%.
- The proposed approach is to a certain extent able to address the challenges of the regulatory domain question answering system. In particular, we address the fact that regulatory documents are usually long by introducing a two-step approach (paragraph and answer selection), also taking advantage of the structure of the documents which usually includes several sections, subsections, and paragraphs.
- From the performed evaluation, the combination of the IR approach for paragraph selection and the MatchLSTM model for answer selection delivers the best results in terms of F1-measure and exact match. In case we consider instead of the top-1 paragraph the top-3 selected paragraphs, the neural network based approach for paragraph selection performs slightly better. In general, both approaches for paragraph selection combined with the answer selection approach deliver comparable results.
- Transfer learning has impact on the domain only if domain-specific training data are taken into consideration. That means: (1) Models trained on other datasets in

order to address the same task do not generalize well on regulatory data. This was observed when SQuAD was used for training and the trained model was tested with in-domain data; (2) However, with careful fine-tuning, the models can substantially increase the performance in the regulatory domain.

5. Related Work

Question Answering over Legal Documents A significant amount of research work has been done in the direction of IR over legal documents. Some examples include the approach introduced by Landhaler et al. [9] which enhances full text search in document collections related to rights and obligations in contracts to find exact and semantically related matches using word2vec. Other approaches combine document embeddings and semantic word measures and natural language processing techniques for retrieving legal case documents [13]. For regulatory documents the efficiency of such techniques has not been tested yet. Also, QA over legal documents is not a new field of research. A QA system for retrieval of Portuguese juridical documents has been proposed by Quaresma et al. [11], able to answer factual questions related to criminal processes. This performs information extraction to build a structured knowledge base from the facts and transforms natural language questions into queries against this knowledge base. However, since only part of the document content can be semantically identified and described many questions remain unanswered. In a different approach, Ranking SVM and Convolutional Neural Networks are employed for building a question answering system able to retrieve answers included in legal articles at paragraph-level [2]. In fact, each article is split into single paragraphs before retrieving answers corresponding to these paragraphs. In our case, we employ an even finer granularity – at text span level. The main disadvantage of the aforementioned systems is that they retrieve relevant documents which may contain the answer rather than the actual answer to a user's question, as we are able to achieve with the current proposed approach.

Reading Comprehension Reading comprehension is a new field of research related to tasks like question answering and machine translation and its goal is "teaching" computers read and understand documents as humans do. To address this problem many deep learning models have been proposed. Wang and Jiang [16] have developed an end-to-end neural architecture for retrieving an answer given a paragraph; this approach is based on match-LSTM model and a Pointer Network [15], a sequence-to-sequence model which allows the generation of answers consisting of multiple tokens (i.e., text span) coming from the original input (i.e., paragraph). The implemented model in the current proposal for answer selection is based on the aforementioned approach. Several other models have been proposed addressing the problem of reading comprehension. One common approach is to use recurrent neural networks (RNNs) in order to predict or generate the answers together with the use of attention mechanisms for matching the words of the question to the corresponding text span in a given passage [5]. Memory Networks [14] have also been applied to reading comprehension. These methods are often slow for long documents because the model needs to be executed sequentially over possibly thousands of tokens. To address the problem of long documents Choi et al. [1] propose to com-

bine a coarse, fast model for selecting relevant sentences and a more expensive RNN for retrieving the answer from those sentences. Similarly, we are following a two-step approach by first selecting the related paragraph and then the corresponding text span.

QA Benchmarks Most of the available datasets for training reading comprehension and QA related models are focusing on the general domain. We have shown in this paper how such datasets can be used for transfer learning to specific domains, in this case the regulatory domain. The Stanford Question Answering Dataset (SQuAD) [12] which was used to train the models in the proposed approach contains more than 100,000 question-answer pairs for documents taken from approximately 500 Wikipedia articles which have been annotated through crowdsourcing. Another large-scale dataset, MS MARCO, consists of 100,000 questions, 1 million passages, and links to over 200,000 Web documents; compared to SQuAD it includes several unanswerable queries and provides human generated answers rather than text spans. SQuAD has been proved more appropriate for the type of problem we are solving, therefore it has been selected for training our models. Apart from SQuAD and MS MARCO, a few other datasets exist addressing the tasks of reading comprehension and open-domain QA like the corpus of cloze style questions from CNN/Daily News summaries [5] and the Children's Book Test [6]. Apart from being too small for deep learning approaches the aforementioned datasets do not map natural questions to text spans directly, thus, they are inappropriate for our approach.

6. Conclusions

In this article we propose a two-step QA system for regulatory documents. We evaluate its various variants and compare it with contemporary IR techniques. Through our experiments, we conclude that the combination of the IR approach for paragraph selection and the MatchLSTM model for answer selection delivers the best results in terms of F1-measure and exact match. Furthermore we find that both transfer learning and data generation mechanisms can significantly improve the performance of the system.

In the future, we aim to further improve the system performance by generating and training our models on larger datasets. We look forward to extensions of this approach over other documents in the regulatory domain, and the generalization of the current system to work across them. Additionally, an interesting direction for further research would be to investigate the model's uncertainity quantification in its predictions. We hope that this would further bolster the adoption of these techniques in the community.

References

[1] E. Choi, D. Hewlett, J. Uszkoreit, I. Polosukhin, A. Lacoste, and J. Berant. Coarse-to-Fine Question Answering for Long Documents. In *Proc. of the 55th Ann. Meeting of the Association for Computational Linguistics, ACL 2017, Vol. 1: Long Papers*, pages 209–220, 2017.

[2] P. Do, H. Nguyen, C. Tran, M. Nguyen, and M. Nguyen. Legal Question Answering using Ranking SVM and Deep Convolutional Neural Network. *CoRR*, abs/1703.05320, 2017.

[3] X. Du, J. Shao, and C. Cardie. Learning to Ask: Neural Question Generation for Reading Comprehension. In *Proc. of the 55th Ann. Meeting of the Association for Computational Linguistics, ACL 2017, Vol. 1: Long Papers*, pages 1342–1352, 2017.

[4] D. A. Ferrucci. Introduction to "This is Watson". *IBM Journal of Research and Development*, 56(3):1, 2012.

[5] K. M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems 28: Ann. Conf. on Neural Information Processing Systems 2015*, pages 1693–1701, 2015.

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

[7] M. Hu, Y. Peng, Z. Huang, N. Yang, M. Zhou, et al. Read+ verify: Machine reading comprehension with unanswerable questions. *arXiv preprint arXiv:1808.05759*, 2018.

[8] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[9] J. Landthaler, B. Waltl, P. Holl, and F. Matthes. Extending Full Text Search for Legal Document Collections Using Word Embeddings. In *Legal Knowledge and Information Systems - JURIX 2016: The Twenty-Ninth Ann. Conf.*, pages 73–82, 2016.

[10] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proc. of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Ann. Conf. on Neural Information Processing Systems (NIPS 2016).*, 2016.

[11] P. Quaresma and I. P. Rodrigues. A Question Answer System for Legal Information Retrieval. In *Legal Knowledge and Information Systems - JURIX 2005: The Eighteenth Ann. Conf. on Legal Knowledge and Information Systems.*, pages 91–100, 2005.

[12] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing, EMNLP 2016.*, pages 2383–2392, 2016.

[13] K. Sugathadasa, B. Ayesha, N. de Silva, A. S. Perera, V. Jayawardana, D. Lakmal, and M. Perera. Legal Document Retrieval using Document Vector Embeddings and Deep Learning. *CoRR*, abs/1805.10685, 2018.

[14] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-To-End Memory Networks. In *Advances in Neural Information Processing Systems 28: Ann. Conf. on Neural Information Processing Systems 2015*, pages 2440–2448, 2015.

[15] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer Networks. In *Advances in Neural Information Processing Systems 28: Ann. Conf. on Neural Information Processing Systems 2015*, pages 2692–2700, 2015.

[16] S. Wang and J. Jiang. Machine Comprehension Using Match-LSTM and Answer Pointer. *CoRR*, abs/1608.07905, 2016.

[17] L. Y. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, and J. Weston. StarSpace: Embed All The Things! In *Proc. of the Thirty-Second AAAI Conf. on Artificial Intelligence, 2018*, 2018.