KAAPA: Knowledge Aware Answers from PDF Analysis

Nicolas Fauceglia, Mustafa Canim, Alfio Gliozzo, Jennifer J Liang, Nancy Xin Ru Wang, Douglas Burdick, Nandana Mihindukulasooriya, Vittorio Castelli, Guy Feigenblat, David Konopnicki, Yannis Katsis, Radu Florian, Yunyao Li, Salim Roukos and Avirup Sil

IBM Research AI

{nicolas.fauceglia, wangnxr, nandana.m, yannis.katsis}@ibm.com,
{mustafa, gliozzo, jjliang, drburdic, vittorio, raduf, yunyaoli, roukos, avi}@us.ibm.com,
{guyf, davidko}@il.ibm.com

Abstract

We present KAAPA (**K**nowledge **A**ware **A**nswers from **P**DF **A**nalysis), an integrated solution for machine reading comprehension over both text and tables extracted from PDFs. KAAPA enables interactive question refinement using facets generated from an automatically induced Knowledge Graph. In addition, it provides a concise summary of the supporting evidence for the provided answers by aggregating information across multiple sources. KAAPA can be applied consistently to any collection of documents in English with *zero* domain adaptation effort. We showcase the use of KAAPA for QA on scientific literature using the COVID-19 Open Research Dataset.

System Overview

End-to-end (E2E) Question Answering (QA) systems (Yang et al. 2019; Chakravarti et al. 2019) based on Machine Reading Comprehension (RC) (Rajpurkar et al. 2016; Yang et al. 2018; Kwiatkowski et al. 2019; Cui et al. 2019; Pan et al. 2019) are effective in providing precise answers to natural language questions. However, they have limitations: **(1)** most QA systems (Yang et al. 2019; Yang, Fang, and Lin 2017) only extract answers from text, while PDF documents are often richer, containing semi-structured information in the form of tables and diagrams; **(2)** precise answers can only be given when questions are specific (Kwiatkowski et al. 2019); frequently, the user's intent is not clear from the initial question, and further exploration is needed to narrow it down; **(3)** supporting evidence for a single answer could be spread across multiple sources, requiring text summarization technology (Feigenblat et al. 2017; Roitman et al. 2020), which is typically not supported by current RC solutions (Yang et al. 2019; Chakravarti et al. 2019).

KAAPA provides a solution for the aforementioned limitations by introducing QA over tables extracted from PDF, Dynamic Facets to narrow down questions and Evidence Summarization in an E2E integrated system. KAAPA consists of two main components: 1) a knowledge induction system, executed *offline* at PDF ingestion time, and 2) an integration of dynamic faceted search, RC and text summarization executed *online* to provide answers to natural language questions. Figure 1 illustrates the architecture.

In the offline component, text is extracted from PDF documents using Smart Document Understanding (SDU), part of IBM Watson Discovery Service[1]. For table extraction, we use our Global Table Extractor (GTE) (Zheng et al. 2020), which leverages specialized object detection models and clustering techniques to extract, for each table, both its bounding box and cell structure. Each document, text and tables, is indexed by an internal search engine[2]. Tables are stored as pseudo-documents containing cell values, surrounding text and intra-document references. KAAPA then automatically induces a Knowledge Graph (KG) from the ingested text. This process builds upon our work presented in (Fauceglia et al. 2019) that provides, among other things: terminology extraction, distributional semantics models and automatic taxonomy induction. All taxonomies can be optionally curated by a Subject Matter Expert using a *Smart Spreadsheet* (Fauceglia et al. 2019).

The online part of KAAPA is in charge of providing summarized answers to the user's questions. Using a passage search strategy, KAAPA retrieves relevant passages. Relevant tables are then identified using a document search strategy on the index of pseudo documents created offline, as described in (Shraga et al. 2020b,a), that achieves state-of-the-art accuracy on table retrieval benchmarks. Dynamic Facets are generated by ranking terms in the induced KG by topic-similarity to the query, and by grouping them by their taxonomy-types, as described in (Mihindukulasooriya et al. 2020). The user can focus the search by selecting a facet, thereby adding the corresponding term as a hard filter to the original query.

The retrieved passages and tables are then fed to the RC system for further analysis. We use GAAMA (Go Ahead Ask Me Anything) (Chakravarti et al. 2019) for this purpose. Given a token sequence ($\mathbf{X}$) consisting of a question and a passage, GAAMA predicts the begin and end of answer spans. Similarly, we use GAAMA to answer natural language questions over these tables, leveraging span selection techniques to identify relevant cells from tables. Specifically, we trained a GAAMA model using the xxlarge v2 version of ALBERT on Open Domain Table QA benchmarks, achieving 0.896 MRR and 85.3% Hit@1 over Wik-

[1] https://www.ibm.com/cloud/watson-discovery

[2] ElasticSearch: https://elastic.co/enterprise-search

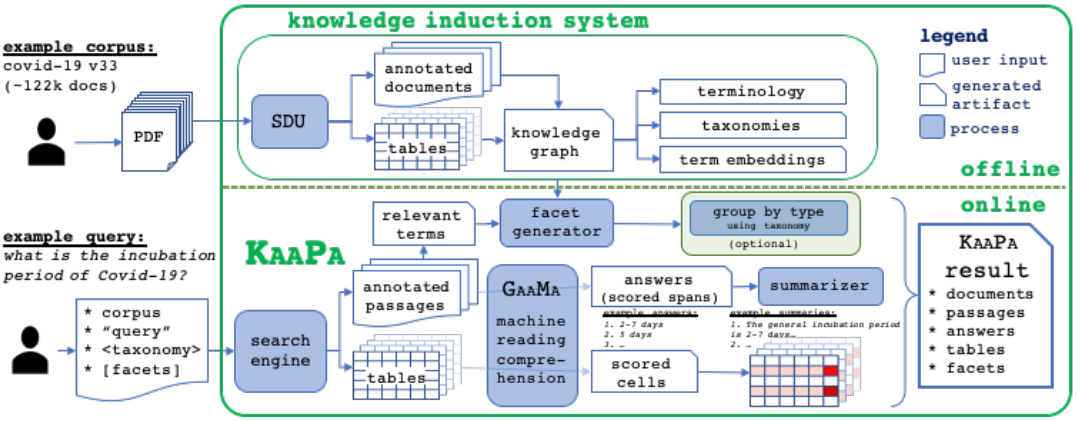

Figure 1: Architecture of the KAAPA system. Showing also an example from the COVID-19 use case.

iSQL benchmark look-up questions. In both table and text QA, the model has been fine-tuned on a collection of open domain QA tasks and no additional fine-tuning is needed on the target domain.

As a final step, KAAPA provides supporting evidence for each returned answer using the *Dual Cascade Cross Entropy Summarizer*, a state-of-the-art unsupervised, query-focused, extractive, multi-document summarizer (Roitman et al. 2020), which receives as input the question, the answer and the concatenation of the passages identified by GAAMA, and returns a summary aggregating and distilling the information found in the supporting passages.

Use Case

To demonstrate KAAPA's zero-shot effectiveness in a new domain, we ingested the COVID-19 Open Research Dataset (CORD-19, v33) (Wang et al. 2020), including PDFs downloaded from links provided in the dataset. The experience of using KAAPA is similar to using a search engine. For example, for the question: *what kinds of complications can occur due to COVID-19?*, KAAPA returns a list of answers (*e.g. acute respiratory distress syndrome, myocardial infarction*). Depending on the user's interest, the search can be further refined using Dynamic Facets. For example, when the user selects the facet *pregnant women*, the search engine will only return documents and answers related to pregnancy-related complications (*e.g. preterm delivery, preeclampsia*). For each returned answer, KAAPA also provides supporting evidence in the form of text snippets containing the surrounding text where the answers have been found. Most of these snippets come from different documents, and express the answer in a diverse way; in order to simplify the user's experience, the system also returns a concise summary of the evidence for each answer. KAAPA also provides the capability to search over tables (Shraga et al. 2020a), returning a ranked list of tables where relevant cells are highlighted using a heat-map approach. A demo video[3] showing the functionalities is provided.

[3]https://ibm.box.com/v/kaapa-aaai21

Evaluation

Qualitative Evaluation. We provided KAAPA after ingesting CORD-19 to a medical doctor (MD) for evaluation. The MD produced 20 queries, covering topics such as transmission, incubation, symptomatology, treatments, and mortality. Evaluation was based on the top 5 returned answers, their supporting evidence and source documents. Overall, KAAPA returned useful concise answers for well-specified queries. For more open-ended queries, the answers returned were less useful but still able to guide the user to relevant passages addressing the underlying information need. For example, for the query *pediatric kawasaki-like disease in COVID-19* where the user's intent was general exploration of this topic, answers included other names this new condition is known by, *e.g. MIS-C, hyperinflammatory shock syndrome*. Dynamic Facets also provided suggestions to focus the search, such as by *clinical presentation* or *intensive care unit*. The MD found KAAPA to be most useful where multiple, sometimes conflicting, answers can be found for a given query. For example, the query *results of remdesivir trial for COVID-19* yielded mixed results from different studies on remdesivir use, with answers such as: *remdesivir was not associated with statistically significant clinical benefits*, and *a significant faster time to recovery but without any difference in mortality*. By presenting all possible answers from the corpus, the user can critically review the literature to develop a more complete and well-informed view of the topic.

Quantitative Evaluation. To evaluate the zero-shot transferability of GAAMA, the underlying QA model, we performed inference on the recent CovidQA (Tang et al. 2020) benchmark. Table 1 shows that GAAMA consistently outperforms the best baselines reported in (Tang et al. 2020) and hence can be regarded as the current state-of-the-art.

	P@1	R@3	MRR
BioBERT + MS MARCO	0.19	0.31	0.31
T5 + MS MARCO	0.28	0.40	0.42
GAAMA	**0.35**	**0.56**	**0.45**

Table 1: GAAMA vs. (Tang et al. 2020) on CovidQA.

References

Chakravarti, R.; Pendus, C.; Sakrajda, A.; Ferritto, A.; Pan, L.; Glass, M.; Castelli, V.; Murdock, J. W.; Florian, R.; Roukos, S.; and Sil, A. 2019. CFO: A Framework for Building Production NLP Systems. *EMNLP-IJCNLP, Demo Track* .

Cui, Y.; Liu, T.; Che, W.; Xiao, L.; Chen, Z.; Ma, W.; Wang, S.; and Hu, G. 2019. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 5882–5888. Association for Computational Linguistics. URL https://doi.org/10.18653/v1/D19-1600.

Fauceglia, N. R.; Gliozzo, A.; Dash, S.; Chowdhury, M. F. M.; and Mihindukulasooriya, N. 2019. Automatic Taxonomy Induction and Expansion. In *EMNLP : System Demonstrations*.

Feigenblat, G.; Roitman, H.; Boni, O.; and Konopnicki, D. 2017. Unsupervised Query-Focused Multi-Document Summarization using the Cross Entropy Method. In Kando, N.; Sakai, T.; Joho, H.; Li, H.; de Vries, A. P.; and White, R. W., eds., *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, 961–964. ACM.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Kelcey, M.; Devlin, J.; Lee, K.; Toutanova, K. N.; Jones, L.; Chang, M.-W.; Dai, A.; Uszkoreit, J.; Le, Q.; and Petrov, S. 2019. Natural Questions: a Benchmark for Question Answering Research. *TACL* URL https://tomkwiat.users.x20web.corp.google.com/papers/natural-questions/main-1455-kwiatkowski.pdf.

Mihindukulasooriya, N.; Mahindru, Ruchi, C.; Md Faisal Mahbub, Deng, Y.; Fauceglia, N.; Rossiello, G.; Dash, S.; and Gliozzo, A. 2020. Dynamic Faceted Search for Technical Support exploiting Induced Knowledge. In *In Proceedings of the 19th International Semantic Web Conference (ISWC)*.

Pan, L.; Chakravarti, R.; Ferritto, A.; Glass, M.; Gliozzo, A.; Roukos, S.; Florian, R.; and Sil, A. 2019. Frustratingly Easy Natural Question Answering. *arXiv preprint arXiv:1909.05286* .

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *EMNLP* URL http://dx.doi.org/10.18653/v1/D16-1264.

Roitman, H.; Feigenblat, G.; Cohen, D.; Boni, O.; and Konopnicki, D. 2020. Unsupervised Dual-Cascade Learning with Pseudo-Feedback Distillation for Query-Focused Extractive Summarization. In Huang, Y.; King, I.; Liu, T.; and van Steen, M., eds., *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, 2577–2584. ACM / IW3C2.

Shraga, R.; Roitman, H.; Feigenblat, G.; and Canim, M. 2020a. Ad Hoc Table Retrieval using Intrinsic and Extrinsic Similarities. In Huang, Y.; King, I.; Liu, T.; and van Steen, M., eds., *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, 2479–2485. ACM / IW3C2.

Shraga, R.; Roitman, H.; Feigenblat, G.; and Canim, M. 2020b. Web Table Retrieval using Multimodal Deep Learning. *SIGIR, 2020* .

Tang, R.; Nogueira, R.; Zhang, E.; Gupta, N.; Cam, P.; Cho, K.; and Lin, J. 2020. Rapidly Bootstrapping a Question Answering Dataset for COVID-19. *arXiv preprint arXiv:2004.11339* .

Wang, L. L.; Lo, K.; Chandrasekhar, Y.; Reas, R.; Yang, J.; Eide, D.; Funk, K.; Kinney, R. M.; Liu, Z.; Merrill, W.; Mooney, P.; Murdick, D. A.; Rishi, D.; Sheehan, J.; Shen, Z.; Stilson, B.; Wade, A. D.; Wang, K.; Wilhelm, C.; Xie, B.; Raymond, D. M.; Weld, D. S.; Etzioni, O.; and Kohlmeier, S. 2020. CORD-19: The Covid-19 Open Research Dataset. *ArXiv* .

Yang, P.; Fang, H.; and Lin, J. 2017. Anserini: Enabling the Use of Lucene for Information Retrieval Research. SIGIR. ACM. ISBN 978-1-4503-5022-8. URL http://doi.acm.org/10.1145/3077136.3080721.

Yang, W.; Xie, Y.; Lin, A.; Li, X.; Tan, L.; Xiong, K.; Li, M.; and Lin, J. 2019. End-to-End Open-Domain Question Answering with BERTserini.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380. Brussels, Belgium: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1259.

Zheng, X.; Burdick, D.; Popa, L.; and Wang, N. X. R. 2020. Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context. *arXiv preprint arXiv:2005.00589* .