Question Answering For Toxicological Information Extraction

Bruno Carlos Luís Ferreira
DEI, CISUC, University of Coimbra, Portugal

Hugo Gonçalo Oliveira
DEI, CISUC, University of Coimbra, Portugal

Hugo Amaro
LIS, Instituto Pedro Nunes, Portugal

Ângela Laranjeiro
Cosmedesk, Coimbra, Portugal

Catarina Silva
DEI, CISUC, University of Coimbra, Portugal

Abstract

Working with large amounts of text data has become hectic and time-consuming. In order to reduce human effort, costs, and make the process more efficient, companies and organizations resort to intelligent algorithms to automate and assist the manual work. This problem is also present in the field of toxicological analysis of chemical substances, where information needs to be searched from multiple documents. That said, we propose an approach that relies on *Question Answering* for acquiring information from unstructured data, in our case, English PDF documents containing information about physicochemical and toxicological properties of chemical substances. Experimental results confirm that our approach achieves promising results which can be applicable in the business scenario, especially if further revised by humans.

2012 ACM Subject Classification Computing methodologies → Information extraction

Keywords and phrases Information Extraction, Question Answering, Transformers, Toxicological Analysis

Digital Object Identifier 10.4230/OASIcs.SLATE.2022.3

Funding This work was partially funded by: the project SafetyDesk: Smart Toxicological Analysis of Chemical Substances (CENTRO-01-0247-FEDER-113485), co-financed by the European Regional Development Fund (FEDER), through Portugal 2020 (PT2020), and by the Regional Operational Programme Centro 2020; and national funds through the FCT – Foundation for Science and Technology, I.P., within the scope of the project CISUC – UID/CEC/00326/2020 and by the European Social Fund, through the Regional Operational Program Centro 2020.

1 Introduction

With the increasing volume of available data, companies need to develop processes for mining information that may be essential for their business. Unfortunately, much of this information is not present in structured databases, but rather in unstructured or semi-structured texts. In many cases, humans are capable of doing this process, however, it can take a long time to complete. *Information Extraction* (IE) emerges as a solution to deal with this problem [4].

In real business scenarios, document processing typically focuses on narrow and specific topics rather than general and wide domains. In such scenarios, annotated data, categorized and labeled for *Artificial Intelligence* (AI) applications, and thus ready for supervised learning approaches, is very limited, and the annotation process is still a challenging task, due to the required time and logistics.

© Bruno Carlos Luís Ferreira, Hugo Gonçalo Oliveira, Hugo Amaro, Ângela Laranjeiro, and Catarina Silva;
licensed under Creative Commons License CC-BY 4.0

In our case study, the problem in question emerged from the necessity of optimising the time it takes to elaborate the toxicological profile of a chemical substance. Currently, the process consists of a human searching for information about the chemical substance and preparing a report with all the relevant information, i.e., physicochemical and toxicological properties. The research resorts to different types of databases, including some where data is structured (e.g., websites) and where data is unstructured (e.g., PDFs, specific papers).

Reviews and reports on toxicological profiles, available in PDF, contain much information written by human experts in an unstructured format, i.e. natural language (English). This includes relevant information, such as physicochemical and toxicological information about substances, which currently needs to be manually extracted for further comparison with the other sources, e.g., for cross validation. This is typically a time-consuming and labor-intensive process.

We address the problem of extracting toxicological information from those PDFs by using state-of-the-art *Transformer* models fine-tuned for Extractive *Question Answering* (QA). The key of our approach is to ask useful questions in order to extract relevant information about toxicological properties given paragraphs from the PDFs.

We next review related work, and describe the task and data preparation. We introduce and elaborate our approach in Section 3, report on some experiments, including preliminary results, in Section 4 and we finally draw conclusions in Section 5.

2 Related Work

Information Extraction from text is an important task of *Natural Language Processing* (NLP), that converts unstructured documents to structured data. There are two main methods for this: rule-based and supervised machine learning. Rule-based methods typically rely on handcrafted textual and linguistic patterns that commonly transmit the entities and relations to extract. In contrast, supervised machine learning exploits features to train a classifier that can distinguish extracted information, either for labelling the sequence of words [13] or for generating entities and relations from it [3].

As supervised machine learning techniques require manually labeled training data, which is one of the major drawbacks of these techniques, unsupervised IE techniques emerged. These techniques extract entity mentions from the text, clusters the similar entities and identify relations [6]. Researchers have introduced Open Information Extraction (OpenIE), an unsupervised machine learning technique, which is a relation-independent paradigm that extracts a large set of relational tuples in an open-domain paradigm [1]. However, given that is an unsupervised method, OpenIE has no idea about the types of entities and relations extracted, so the usage of other knowledge bases from external sources is necessary in order to learn the relations in a corpus [1], a drawback of OpenIE.

Great advances taken in the state of the art in 2017 with the introduction of the *Transformer* neural networks [14] and the consequent emergence of *neural language models* (LMs), like BERT [5], RoBERTa [9], ELECTRA [2], which can be fine-tuned for a broad range of NLP tasks.

An alternative to the previous approaches for IE, especially when lacking training data, is to formulate IE as a Question Answering problem, using transformer models fine-tuned for this task. In order to do so, other researchers [11, 10, 7] acquired a pre-defined list of required information and represented it as key phrases, *e.g.,* “Name of institution” or “Deadline for bidding” [10]. Using the pre-defined key phrases, they can be considered as a question, or part of one, the input document can be treated as the context and the extracted information from a document can be considered as an answer.

3 Our Approach

Within the broad space of business documents, as mentioned in Section 1, we were faced with the challenge of accelerating the process of filling toxicological reports, so we focused on one specific type of documents: studies of individual chemical substances. Our objective is to extract specific properties (information) from those studies (input documents).

However, given an input document, there are multiple pages, multiple paragraphs and multiple phrases that contain information, regarding physicochemical and toxicological properties of chemical substances. Simply providing the complete document to the LMs is a problem because the LMs used are limited in the size of the context. To tackle that problem we decided to do divide the documents into sections, where each section contains the information regarding a specific property of the chemical substance.

Based in our objective and restriction our approach follows two main steps:

1. Identify the section of the document about a specific property.
2. Get the information for that property by asking a question to the model, using the previously identified section as context.

3.1 Information Pre-processing

The pre-processing process consists of dividing the input document into sections, where each section contains the information regarding a specific property of the chemical substance. That way, we minimize the context given to the LMs, eliminating noise, i.e. parts of the document not relevant for each property.

As a visual example, in Figure 1 we have the properties A and B where the property A has information in section *Blue* and *Red* and property B has information in section *Yellow*. It is not necessary to search for information regarding property B in all the three sections, only in section *Yellow*.

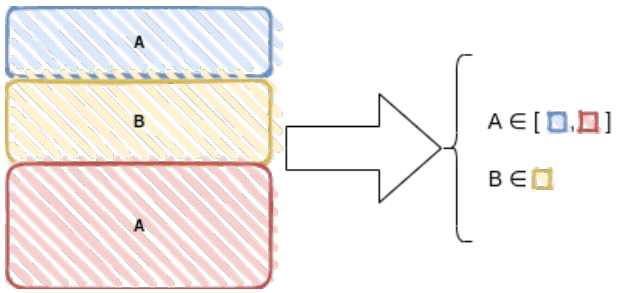


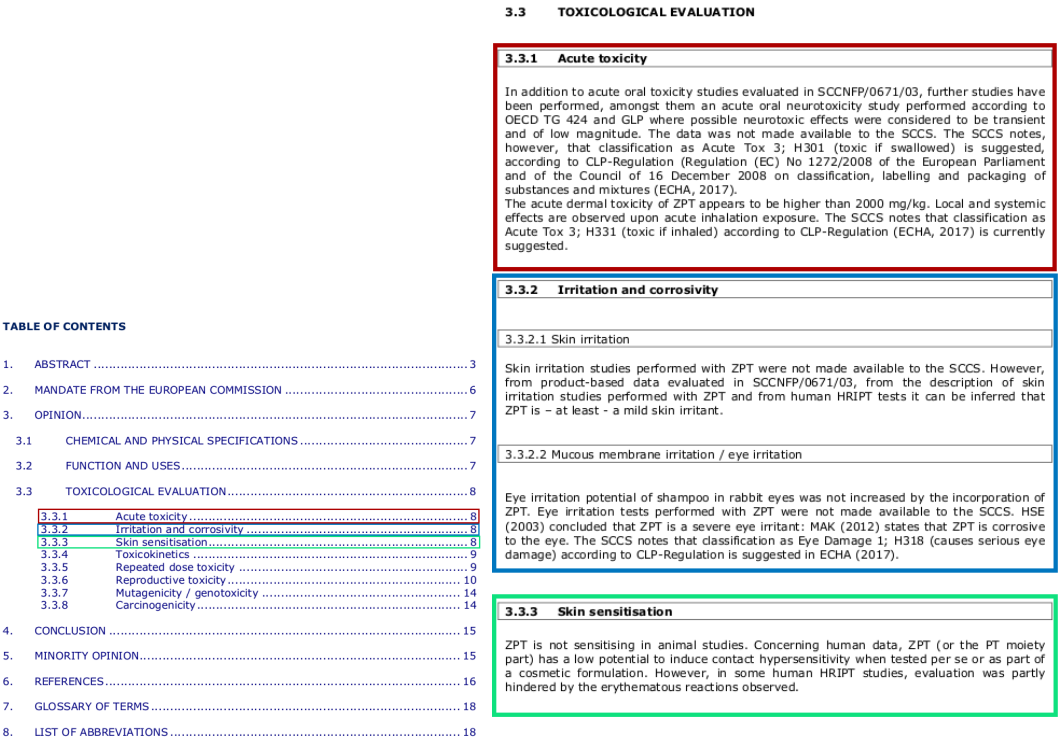
Figure 1 Graphical example of the pre-processing process.

In our approach, the division of the document in sections is based on the *Table of Contents* (TOC) that the document has. This process is similar to that of a human when navigating and searching the document using the Index/TOC. It is possible to use the TOC as the reference point for the division of the document into sections because one is commonly present in the reports of toxicological profiles we have been using.

The input PDF documents were converted to text with the *pdfplumber*[1] parser, and, combined with Regular Expressions, we could obtain the TOC of the document. The usage of the TOC allows us to find the start and the end of each section, i.e., by considering the

[1] `https://github.com/jsvine/pdfplumber`

number and title of the sections, where the start corresponds to the section title obtained in the obtained TOC and the end of the section is the starting of the next section with the same hierarchical level. Figure 2 represents the pre-processing process, where the information obtained from the TOC (Figure 2a) help us divide the document into sections (Figure 2b).



(a) Table of contents of document with sections visually indicated. **(b)** Page of document with sections visually divided.

Figure 2 Example of the pre-processing process in a document.

Being able to divide the input document into sections allows us to increase the performance and reduces the noise because, by limiting the context that we provide to the LMs, we can guarantee that the answers obtained are connected to the substance's property information.

3.2 QA for IE

Having the context defined, we can use the QA models for extracting information. In order to do so, we need to identify and set questions related to the context and to the information that we want to obtain. For this task, we can explore available Transformer models fine-tuned in the *The Stanford Question Answering Dataset* [12] (SQuAD), which includes paragraphs (contexts), questions about them, and extracted answers. SQuAD has two versions, 1.1 and 2.0. The main difference between them is that version 1.1 contains 100,000+ question-answer pairs on 500+ contexts, while version 2.0 combines the 100,000 questions in SQuAD 1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. This means that models trained with SQuAD 2.0 not only answer questions when possible, but can also determine when no answer is possible from the given paragraph and abstain from answering [12].

As SQuAD uses the Six W's (Who, What, When, Where, Why and How) in the formulation of the questions, we also need to create questions of this kind, regarding each information that we want to extract. For example, in the sentence present in one of the PDF documents used, "Eye irritation potential of shampoo in rabbit eyes was not increased by the incorporation of ZPT" we want to obtain the species that the test applies to, so we can formulate a question as "What is the species?". Given the sentence (as the context) and the question, we hope to obtain from the QA models the right answer, in this case, "rabbit".

After the pre-processing step, we can give to the QA models each section of the document that corresponds to a specific property of a substance as context. In order to optimize our approach, we need to create the right set of questions per section.

4 Experiments

We focus our experiments in one source of studies of individual chemical substances: *Scientific Committee on Consumer Safety* (SCCS) Opinions[2]. For experimentation purposes, we focus on a subset of 60 documents, issued by the Committee[3], dated from April 2016 to December 2021. Our objective is to extract relevant information about certain toxicological properties of substances. Table 1 shows a sample of relevant information about the target toxicological properties for this case study.

Table 1 Substances properties information.

Substance Property	Information to Extract
Repeated Dose Toxicity	NOAEL[4] value
Acute Toxicity	Species used in study; OECD[5] Guideline used; Exposure route
Irritation	Species used in study; OECD Guideline used; Exposure route
Mutagenicity	OECD Guideline used; Classification
Skin Sensitization	OECD Guideline used; Classification; Concentration used in study
Carcinogenicity	Species used in study; OECD Guideline used; Classification
Photo-induced Toxicity	OECD Guideline used; Classification
Reproductive Toxicity	Species used in study; OECD Guideline used; Classification

4.1 Setup

The source documents were pre-processed (see Section 3.1) and a set of questions was formulated for each information to extract (see Table 2). In this case study, all the questions start with "what" because, by trial and error, we noticed that relevant information was frequently obtained with questions like "what is the *[specific information to extract]* ?".

In our experiments we tested the set of questions in three different QA models, all available from the Huggingface Transformers hub, and usable from the transformers library[6]: *BERT-base-cased-squad2*[7], *RoBERTa-base-squad2*[8] and *BioBERT-v1.1-pubmed-squad-v2*[9].

[2] `https://ec.europa.eu/health/scientific-committees/scientific-committee-consumer-safety-sccs/sccs-opinions_en`
[3] `https://ec.europa.eu/health/scientific-committees/former-scientific-committees/scientific-committee-consumer-safety-2016-2021/sccs-opinions-2016-2021_en`
[4] No Observed Adverse Effect Level
[5] Organisation for Economic Co-operation and Development
[6] `https://huggingface.co/`
[7] `https://huggingface.co/deepset/bert-base-cased-squad2`
[8] `https://huggingface.co/deepset/roberta-base-squad2`
[9] `https://huggingface.co/ktrapeznikov/biobert_v1.1_pubmed_squad_v2`

Table 2 Example of set of questions per property tested.

Substance Property	Questions
Repeated Dose Toxicity	What is the NOAEL value?
Acute Toxicity	What is the guideline?;What is the species?
Irritation	What is the guideline?;What is the species?
Mutagenicity	What is the Guideline?;What is the conclusion?
Skin Sensitization	What is the Guideline?;What is the conclusion?;What is the concentration?
Carcinogenicity	What is the species?;What is the Guideline?;What is the conclusion?
Photo-induced Toxicity	What is the Guideline?;What is the conclusion?
Reproductive Toxicity	What is the Guideline?;What is the species?;What is the conclusion?

Although all the models are Transformers, they also have their differences, at the architecture level or at the pre-train level, that have impact in the results. BERT is the basic model fine-tuned for QA. When released, it achieved state-of-the-art performance in many NLP tasks, including QA [10]. RoBERTa builds on BERT and modifies key hyperparameters, removing the next-sentence pre-training objective and training with much larger mini-batches and learning rates[10]. BioBERT is a domain-specific language representation model, pre-trained on large-scale biomedical corpora that, while having the same architecture, outperforms BERT in a variety of biomedical text mining tasks [8].

We provided the QA models with:

1. The sections of the document that contain information about each specific substance property as context.
2. The set of questions that we defined specifically for each substance property.

As a result, we expected to extract relevant information about each substance property.

4.2 Evaluation metrics

In order to achieve a quantitative evaluation of our experiments, a confusion matrix was built with the number of *True Positives* (TP), *False Positives* (FP), *False Negatives* (FN) and *True Negatives* (TN). In the context of this work, those are defined as:

- TP: There is information in the document to be extracted and the information extracted is correct;
- FP: There is no information in the document to be extracted but there is some information extracted or there is information in the document to be extracted but the information extracted is not correct;
- FN: There is information in the document to be extracted but there is no information extracted;
- TN: There is no information in the document to be extracted and there is no information extracted;

The extracted outputs were matched to ground-truth data, i.e. the extracted information was manually compared with the information present in each document tested. Using the outcomes, we calculated the precision, recall, and F1 score in order to evaluate the performance of our experiments.

[10] `https://huggingface.co/docs/transformers/model_doc/roberta` (March 2022)

4.3 Combination Process

We can use the models individually or, to take full advantage of them, we can aggregate the answers obtained from each. The combination process consists of using an answer, i.e. information extracted, if the same or similar answer was given as an output from at least two of three models, as shown in the example of Table 3. At the point of this evaluation the combination process was done manually despite the development of the process having already started using text similarity measures.

Table 3 Visual example of combination process.

Original text excerpt

```
Guideline:  OECD TG 429 Skin Sensitization:  Local Lymph Node Assay 24 th April 2002
Species/strain:  female CBA/J mice
Group size:  4 mice per group, 20 animals per experiment, 2 independent experiments
Batch: RO060245B 002 L 002
Concentration: 0.1, 1 and 10 %
Study period: 13 June - 12 September 2008
The test item was not soluble in any of the recommended vehicles.  However, a homogeneous suspension was
obtained at the maximum tested concentrations of 10% and 15%, with propylene glycol, after sonication for
10 minutes.  Therefore propylene glycol was selected as vehicle.  On days 1, 2 and 3 of each experiment, a
dose-volume of 25 µL of the control or dosage form preparations was applied to the dorsal surface of both ears.
On day 6 of each experiment, all animals of all groups received a single intravenous injection of 20 µCi of 3H-TdR.
SCCS comment
Based on this LLNA study in which a maximum concentration of 15% was used, A164 is considered not to
have skin sensitising potential.
```

	What is the Guideline?	What is the conclusion?	What is the concentration?
BERT	OECD TG 429		0. 1, 1 and 10 %
BioBERT	Skin Sensitization : Local Lymph Node Assay	Skin Sensitization : Local Lymph Node Assay	0. 1, 1 and 10 %
		not to have skin sensitising potential	25 µL
RoBERTa	OECD TG 429	The test item was not soluble in any of the recommended vehicles	
		A164 is considered not to have skin sensitising potential	
Combo	OECD TG 429	A164 is considered not to have skin sensitising potential	0. 1, 1 and 10 %

4.4 Results and Discussion

From the 60 documents gathered, 10 were randomly selected for testing our approach. The pre-processing process worked as planned and we were able to find each section regarding each property in the documents without faults in the process. From the 10 documents, the average section size was 899 tokens, the minimum size was 27 and the maximum of 4860 tokens.

We report the performance of each QA model in Table 4, individually, and in Table 5, after the combination process. Both tables also include the micro average, where all the outcomes are taken into account, in order to deliver a fair general evaluation of model performances. For this evaluation the information extracted was not verified by the expert that currently gathers the information manually despite direct contact throughout the development.

By first analysing each QA model individually (Table 4) we were able to understand that some optimizations can be developed even though some strong results were obtained. In some cases the precision and the recall were perfect, which can be due to the disposition of the information in the document, i.e., better results can be achieved if the information is present in bullet points than if it is in the middle of the sentences. In terms of optimisation, we used the same set of questions for each model and there are some performance gains if we use each model in its strong points. For example, *BioBERT*, pre-trained in biomedical data,

Table 4 Individual evaluation of QA models on SCCS documents.

	BERT			BioBERT			RoBERTa		
	F1	P	R	F1	P	R	F1	P	R
Acute Toxicity Information	0.77	0.87	0.70	0.74	0.59	1.00	1.00	1.00	1.00
Irritation Information	0.68	0.76	0.61	0.76	0.61	1.00	0.85	0.85	0.85
Skin Sensitisation Information	0.86	0.82	0.92	0.72	0.56	1.00	0.85	0.84	0.87
Mutagenicity Information	0.67	0.53	0.92	0.57	0.40	1.00	0.71	0.55	1.00
Carcinogenicity Information	0.84	0.78	0.91	0.66	0.50	1.00	0.58	0.43	0.90
Photo-induced Toxicity Information	0.44	0.40	0.50	0.58	0.41	1.00	0.54	0.37	1.00
Reproductive Toxicity Information	0.80	0.66	1.00	0.70	0.54	1.00	0.73	0.57	1.00
Repeated Dose Toxicity Information	1.00	1.00	1.00	0.80	0.66	1.00	1.00	1.00	1.00
Micro Average	0.76	0.68	0.85	0.64	0.48	1.00	0.76	0.65	0.94

Table 5 Approach evaluation on SCCS documents.

	BERT + BioBERT + RoBERTa		
	F1	Precision	Recall
Acute Toxicity Information	1.00	1.00	1.00
Irritation Information	0.89	0.96	0.83
Skin Sensitisation Information	0.96	0.96	0.96
Mutagenicity Information	0.78	0.65	0.96
Carcinogenicity Information	0.84	0.80	0.88
Photo-induced Toxicity Information	0.75	0.60	1.00
Reproductive Toxicity Information	0.85	0.74	1.00
Repeated Dose Toxicity Information	1.00	1.00	1.00
Micro Average	0.87	0.82	0.93

was the best model for acquiring the names of species, but the worst for identifying guidelines. Even in terms of time, there are gains if we just use the *BioBERT* for identification of the species and nothing else.

Also regarding evaluation, the definition of each outcome, i.e. TP or FP, is subjective for some information Some can be a quantitative value (e.g.,"357 mg/kg bw/day") or a species name (e.g., "Rat / F344 / DuCrlCrlj"), where it is easy to define if the extracted value is a TP or a FP, but sometimes, the extracted information is a short sentence (e.g., "There was no evidence of carcinogenic activity"), which is subject to human interpretation.

By analysing Table 5, we confirm that there are gains when the models are combined, both in terms of precision and recall, when compared with the individual models results. In general, and despite the limited set of questions, we can affirm our approach obtained solid results. Still, in the future, we can experiment with more and different questions, in order to achieve better performances.

Overall, we would characterize our approach and experiments as an important step into extracting information from unstructured documents. In our point of view, this means that, although the human cannot be replaced, our approach can supply them with a set of extracted information that the they can: (i) accept as is, (ii) change or complement minimally or (iii) use as the starting point of a more thorough search.

5 Conclusions

In this paper we treat *Information Extraction* as a *Question Answering* problem and propose an approach that, with limited data, can be a solution for the former. To do that, we take advantage of state-of-the-art extractive QA models. We conducted experiments and analysis on one source of studies of individual chemical substances, and the results obtained are promising, although performance gains can be achieved with some optimizations, in particular, achieving the right set of questions to use with each QA model.

In the future, we will work on generalizing this approach for documents where the TOC is not available, which will enable its application to different sources of information on chemical substances (other than SCCS); on automating the combination and evaluation processes, where text similarity measures like ROUGE can be considered; and involve the expert in the final evaluation of the results. We plan to make this approach available through a REST API, which will provide an easier integration with applications like cosmetics regulatory software, and also consider the application of this approach to other domains.

References

1 Sally Ali, Hamdy Mousa, and M Hussien. A review of open information extraction techniques. *IJCI. International Journal of Computers and Information*, 6(1):20–28, 2019.

2 Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint*, 2020. `arXiv:2003.10555`.

3 Lei Cui, Furu Wei, and Ming Zhou. Neural Open Information Extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 407–413, 2018.

4 A. Cvitaš. Information extraction in business intelligence systems. In *The 33rd International Convention MIPRO*, pages 1278–1282, 2010.

5 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol 1 (Long and Short Papers)*, pages 4171–4186. ACL, 2019.

6 Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1535–1545, 2011.

7 Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. A question answering approach to emotion cause extraction. *arXiv preprint*, 2017. `arXiv:1708.05482`.

8 Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

9 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*, 2019. `arXiv:1907.11692`.

10 Minh-Tien Nguyen, Dung Tien Le, and Linh Le. Transformers-based information extraction with limited data for domain-specific business documents. *Engineering Applications of Artificial Intelligence*, 97:104100, 2021.

11 Minh-Tien Nguyen, Dung Tien Le, Nguyen Hong Son, Bui Cong Minh, Akira Shojiguchi, et al. Information extraction of domain-specific business documents with limited data. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

12 Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of 56th Annual Meeting of the Association for Computational Linguistics (Vol 2: Short Papers)*, pages 784–789. ACL, 2018.

13 Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. Supervised Open Information Extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, 2018.

14 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.