

A.4: Final Project - Creating Chatbot: Walter Chatbot

Melek Mizher[†]

December 4, 2022

[†] Address to which correspondence should be addressed:

melekmizher@me.com

1.0 Introduction and Problem Statement

The objective of this study is to provide a useful framework for the development of a general study aid for all of the chemical elements. We propose a mechanism to allow for students to ask questions related to any of the chemical elements or for a summary of general information about them. This is designed to provide quick answers to user questions to allow for the resolution of quick and simple general questions about any of the currently described 118 elements. This can be useful to provide quick guidance and information to students that are learning basic information about chemical elements for them to have an general grasp of knowledge about those elements. This is done by providing a chatbot that answers student questions and provides students with information based on extracts from wikipedia related to the elements. WALTER(R) stands for Word Analysis Leveraging Tendencies of Elements for Restricted Reasoning.

2.0 Literature Review

Effectively providing quick answers to questions from students can allow for better results that can provide for improved student outcomes as seen in Wen, Zhu, and Zhang (2022). They propose a system through which intelligent tutoring allows for an increased of exam scores for students that pursue a self-regulated learning style, therefore similar in concept to online courses. Average scores for a 100-question exam were about 76% for the self-instructed group, ranged from 84% to 82% in groups with traditional instruction based on class size, and retained an average of about 81.3% for the proposed method using a chatbot regardless of class size. While there is still an advantage for traditional classes, a self-instructed course could benefit from this type of study aid. Carlander-Reuterfelt et al. (2020) provide a general framework for design of an efficient chatbot that teaches concepts

about Data Science. They identified three important pedagogical aspects of the design process. These include the direct definitions of a concept, the use of examples through analogy that provide learner support, and the human need to satisfy some small-talk and personalize the communication. Zheng et al. (2018) propose the usage of a method to automatically index information to separate the data set so that a question answering model can better hone into the important information to provide better answers for more complex questions. They also provide strategies to perform entity-level and structure-level disambiguation.

3.0 Data

The data used in this experiment was gathered from the wikipedia articles for all 118 chemical elements using the Wikipedia API in python. The important text parts of the article remain and are used to generate summaries for the elements that are always four sentences of length. Element names and symbols can both be used to identify the user question and connect it to the appropriate data for the model to work upon. The built corpus is then stripped from punctuation, all characters (including user inputs) are processed to lowercase and any special tags and characters are removed from the data using regex.

4.0 Research Design and Modeling Methods

The design compares the usage of multiple preexisting models to identify which one provides the most useful result with regards to a set of questions. The models used were all based on question answering through different processes. The models studied and the reasons for which they were chosen include the following:

Model Name	Description
distilbert-base-cased-distilled-squad	BERT-based “distilled” model that retains 95% of performance with much faster time.
deepset/roberta-base-squad2	Base model that learns representations from English language in a self-supervised form.
deepset/tinyroberta-squad2	Roberta model with a smaller training dataset.
deepset/minilm-uncased-squad2	NLP model that uses self-attention for embedding and compression of arguments.
mrm8488/longformer-base-4096-finetuned-squadv2	Finetuned for Long documents. Uses local and global attention. Task-specific focus.

Table 1: The five different models used in this experiment and their main attributes for which they were selected.

The structure of the chatbot gives the user a welcome message and asks for the user’s name. The chatbot then proceeds to identify the element of interest for the user, confirms the element, and provides an opportunity for the user to then ask their question. It then takes a few seconds for the program to return an answer to the user’s question. In the example below, such an example of the conversation with the AI chatbot is displayed.

```

Melek: Carbon

Walter: Melek, I see that you want to learn more about Carbon, is that correct?

Melek: Yes

Walter: Got it, Melek!, What would you like to learn about Carbon?

Melek: Is it important to life?

Walter: Give me a few seconds to think about it.

Downloading: 0% | 0.00/3.64k [00:00<?, ?B/s]

Walter: carbon is essential to all known living systems

```

Figure 1: User conversation querying Walter about the chemical element Carbon.

In this example, Walter was able to identify and provide a valid answer from the dataset contained in the data after it was ran through the pretrained model. The Walter chatbot

can also provide summaries a chemical element of interest.

5.0 Results

The results consist of two different categories. The first is the generation of summaries for a chemical element of interest and the second part is the generation of answers based on a user question about a specific chemical element. For the generation of summaries, the algorithm uses the wikipedia article for the chemical element of interest using the wikipedia API to provide a simple and concise four-sentence long summary. This is a simple and straightforward mechanism to extract a summary from the wikipedia article by taking advantage of the functionality of their API.

Walter: Got it, Melek!, What would you like to learn about Hydrogen?

Melek: I would like to have a summary about Hydrogen

Walter: Hydrogen is the chemical element with the symbol H and atomic number 1. Hydrogen is the lightest element. At standard conditions hydrogen is a gas of diatomic molecules having the formula H₂. It is colorless, odorless, tasteless, non-toxic, and highly combustible.

Walter: Got it, Melek!, What would you like to learn about Titanium?

Melek: I want a summary about it

Walter: Titanium is a chemical element with the symbol Ti and atomic number 22. Found in nature only as an oxide, it can be reduced to produce a lustrous transition metal with a silver color, low density, and high strength, resistant to corrosion in sea water, aqua regia, and chlorine.

Titanium was discovered in Cornwall, Great Britain, by William Gregor in 1791 and was named by Martin Heinrich Klaproth after the Titans of Greek mythology. The element occurs within a number of minerals, principally rutile and ilmenite, which are widely distributed in the Earth's crust and lithosphere; it is found in almost all living things, as well as bodies of water, rocks, and soils.

Hydrogen

Titanium

Figure 2: User conversation querying Walter for summaries about the chemical elements Hydrogen and Titanium.

Other user queries that do not ask about a summary can generate more specific answers from Walter. The previously discussed five pretrained models were tested to understand their suitability and functionality for the problem of interest. Six basic questions with proper spelling and grammar were made to generate answers from the models.

Model Question	distilbert-base-cased-distilled-squad	deepset/roberta-base-squad2	deepset/tinybert-squad2	deepset/minilm-uncased-squad2	mrm8488/longformer-base-4096-finetuned-squadv2
What are the uses of Hydrogen?	fossil fuel processing	acid base reactions	when hydrogen forms a compound with a more electropositive element	when hydrogen forms a compound with a more electropositive element	hydrogen lifted airships were used as observation platforms and bombers
Is Carbon Important to life?	carbon is essential to all known living systems	carbon is essential to all known living systems	carbon is essential to all known living systems	carbon is essential to all known living systems	carbon occurs in all known organic life
Iron is what type of element?	telluric iron	chemical	cell contains free iron	comparably soft and ductile and easily cold forged	free iron
What type of element is Iron?	anionic complexes	chemical	cell contains free iron	iron is the most abundant element on earth	free iron
What can you make with Lithium?	multi center bonding	lithium ion batteries	fat produces a soap made of lithium stearate	both stable isotopes of lithium can be laser cooled and were used to produce	lithium carbonate but also releases oxygen
What are the dangers of Uranium?	health hazards	chemical poisoning by uranium oxide	health hazards	chemical poisoning by uranium oxide rather than radioactivity	health hazards

Table 1: Comparison of the answers across the five different models of interest. Shades of green show correct answers (dark green) and correct/slightly irrelevant answers (light green). Yellow displays answers that although they may be correct, they present no useful/relevant information to the user. Answers marked in red are either completely irrelevant or incorrect.

6.0 Analysis and Interpretation

From the results in Table 1, different models were able to come up with correct answers. Although some of which were rather irrelevant to the question even if they were correct in a literal sense. It is important to note that although some of the models may have had a smaller pool of training data, they could still come up with relevant and insightful answers that sometimes surpassed those of the more complete models. For example, tinyroberta identified the production of lithium stearate from lithium as a soap. It is correct to mention that it is a soap since that is part of its attributes, however, this may create confusion in a user as it does not specify that it is not used as a soap in the regular sense of the word, but rather as a high-temperature industrial lubricant. The model that tended to produce the most concise and insightful answer was the base RoBERTa model in every tested case. The larger training data that was used provides for a reasonable query processing time based on the quality of the answer. Simpler models could not come close in producing quality answers. The distilbert model provided a great answer for the first question as it provided a direct use of hydrogen which can indeed be used for fossil fuel processing. In contrast, the base RoBERTa model suggested acid base reactions as the answer. It is correct to consider these reactions as requiring hydrogen to occur, however, these are not a “use” of hydrogen in a literal sense. In the process of developing a chatbot, it is important to consider the logic behind the words and how humans will understand the answers from the model. In the question regarding the dangers of uranium, the MiniLM model added extra words compared to the RoBERTa model that if used correctly could have added more information about the impact of radiation with uranium. However, these words are better left unsaid in this case as they could cause confusion due to the lack of proper syntax.

7.0 Conclusion

Based on the tested information, the RoBERTa model (deepset/roberta-base-squad2) serves as the most useful of all the models as it was able to provide a correct or valid answer for all cases in the tested set of questions. The usage of a purely TF-IDF process to generate answers from the dataset would not have been sufficient and it was crucial to use pretrained models based upon larger datasets of natural language so that the model could generate appropriate and logical answers from the text. However, the process only uses the wikipedia article as the single source of data for the specified element. Therefore there may not be enough information for some of the elements and this may prevent simpler and faster algorithms from generating good answers from the text. It could also allow for the more complete models to provide more insightful answers. The greatest disadvantage is the disconnection to a knowledge graph as that would allow for more specific numerical questions to be asked.

8.0 Future Work

Improvements could be made to the structure of this chatbot by incorporating an ontology such as the Ontology of Chemical Elements (OCE) which was initially developed as a community effort in 2020. Using an ontology could provide with specific and correct answers to questions that asked for numerical information, such as “what is the atomic mass of Helium?” and for comparative questions. However, the scope of the project was just descriptive and specific to each of the elements. To expand on the capacity of the chatbot, the most logical step would be to incorporate ontology-based question answering that provides a level of flexibility to discriminate and generate logical answers for user questions. Other improvements that can be made is to increase the dataset through providing a large collection of different articles for each of the elements together with question-answer pairs for each to

provide for a solid foundation that can allow the models to perform at a greater level. The chatbot could also be adapted so that it is friendlier to user misspellings or other errors.

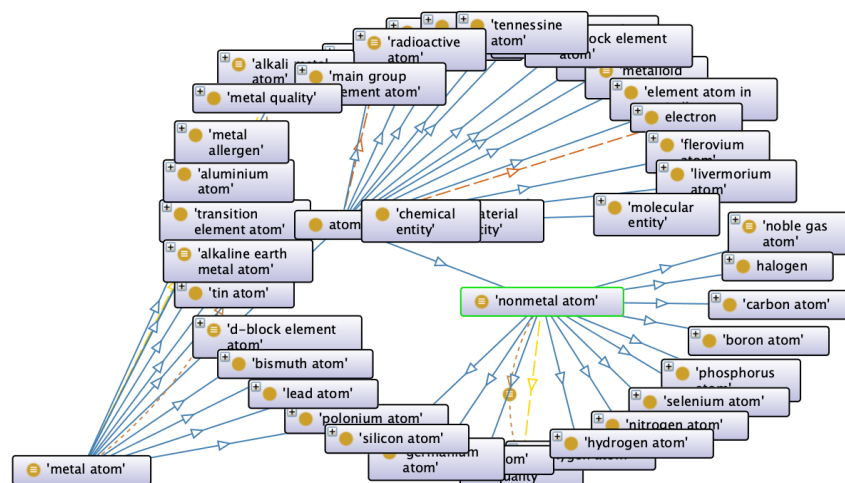


Figure 3: Sample from the Ontology of Chemical Elements emphasizing the categorization of atoms, specifically their division in between metal and nonmetal atoms..

Bibliography

- Carlander-Reuterfelt, Daniel, Alvaro Carrera, Carlos A. Iglesias, Oscar Araque, Juan Fernando Sanchez Rada, and Sergio Munoz. "Jaicob: A Data Science Chatbot." IEEE Access 8 (2020): 180672–80. <https://doi.org/10.1109/access.2020.3024795>.
- "Ontology of Chemical Elements: NCBO Bioportal." Ontology of Chemical Elements | NCBO BioPortal. Accessed December 4, 2022. <https://bioportal.bioontology.org/ontologies/OCE>.
- Wen, Yu, Xinhua Zhu, and Lanfang Zhang. "CQACD: A Concept Question-Answering System for Intelligent Tutoring Using a Domain Ontology with Rich Semantics." IEEE Access 10 (2022): 67247–61. <https://doi.org/10.1109/access.2022.3185400>.
- Zheng, Weiguo, Jeffrey Xu Yu, Lei Zou, and Hong Cheng. "Question Answering over Knowledge Graphs." Proceedings of the VLDB Endowment 11, no. 11 (2018): 1373–86. <https://doi.org/10.14778/3236187.3236192>.