# Using Bloom's Taxonomy to Classify Question Complexity

**Sabine Ullrich**
Research Institute CODE
Universität der Bundeswehr München
sabine.ullrich@unibw.de

**Michaela Geierhos**
Research Institute CODE
Universität der Bundeswehr München
michaela.geierhos@unibw.de

## Abstract

Question answering is widespread and a variety of answer taxonomies exists in research that divides responses into simple and complex. Multi-hop answering has become popular when the complexity of questions and answers increases. However, determining when multi-hop reasoning becomes necessary is not yet clear.

We propose to apply Bloom's taxonomy to the determination of question complexity in question-answering systems. Originating in pedagogy, Bloom's taxonomy measures question complexity to determine learning progress levels. Subsequently, the determined question complexity can help in deciding whether an entity or phrase is sufficient as an answer or whether reasoning chains should be given.

## 1 Introduction

When determining the answer type in a question-answering (QA) system, the question type must be considered first. While entities or short sentences are sufficient for simple, factual questions, more complex questions require more complex answers. For example, simple product-related questions, such as "Does Kindle support Japanese?", can be easily answered by a yes/no response. When extracting interpretative questions that require logical thinking, reasoning chains can be used to generate answers. Imagine a complex question such as "What is the current situation in Syria?". Answering this question is not easy and cannot be done by a simple knowledge graph or ontology. To explain why this answer is correct and to provide a cohesive line of argumentation, multi-hop reasoning chains are required to connect successive propositions.

While several approaches exist that present taxonomies for question and answer types, the complexity of questions has not yet been measured to classify the required answers. Assuming that complex questions require complex answers, we need to ask the question "What makes a question complex?". How can we determine the complexity of a question and at what level of complexity are multi-hop reasoning chains useful or even essential?

In pedagogy, Bloom's Taxonomy of Educational Objectives (Bloom et al., 1956) helps to capture a learner's level of understanding. At the lowest level of the taxonomy, simple memorization is required to reproduce a fact or concept, while as the level increases, the abstraction level also increases. The lower levels serve as base knowledge, while higher levels represent the deeply processed knowledge that can be abstracted and transferred for specific purposes (Cannon and Feinstein, 2005).

This paper examines how Bloom's Taxonomy can be used to classify questions in QA systems according to their complexity. Furthermore, it discusses which factors contribute to the complexity of a question and when multi-hop reasoning is required instead of simple information extraction.

## 2 Related Work

Several approaches attempt to classify answers in QA systems by constructing a question taxonomy. Questions are grouped either flatly (Eichmann and Srinivasan, 1999; Litkowski, 1999) or hierarchically (Takaki, 2000; Suzuki et al., 2003). Kim (2014) proposes a method for defining answers and ambiguity within questions. Moreover, taxonomies exist for specific question types such as the taxonomy for opinion questions (Bayoudhi et al., 2013), classifications based on data source, analysis types, and response forms (Mishra and Jain, 2016).[1] However, none of these surveys defines how these taxonomies can be used to calculate question complexity.

[1] For an extensive list see Sundblad (2007).

Datasets containing multi-hop reasoning chains are widely used (Yang et al., 2018; Jhamtani and Clark, 2020; Wiegreffe and Marasović, 2021). Reasoning chains provide appropriate answers to questions posed in the respective datasets. For general questions asked in QA scenarios, it is unclear if or when a multi-hop reasoning chain is required as an answer. This is because question complexity measurement and reasoning chains have not yet been combined.

Often, question complexity is used in education to determine the difficulty of student exams. For example, Luger and Bowles (2013) measure the difficulty of multiple choice questions. Research on community QA services is often domain-specific, comparing the difficulty of topic-related words within certain domains (Liu et al., 2013; Wang et al., 2014). Others use provided meta-information such as user expertise to estimate question difficulty (Sun et al., 2018) or measure relative complexity by comparing users' questions (Thukral et al., 2019).

Research most closely related to ours comes from Padó (2017), which shows how Bloom's Taxonomy can approximate the difficulty of questions in a short-answer corpus. Together with measuring the diversity of student responses, the difficulty can be estimated from lower to higher levels of the taxonomy. In addition, textual entailment methods can infer levels from the question wording (Anderson and Krathwohl, 2014). However, their approach is only used in the context of grading students, so we propose to adapt it for measuring question complexity in QA systems.

## 3 Approach

Our method combines Bloom's Taxonomy (Bloom et al., 1956) and question classification for QA systems. We plan to classify the difficulty of questions by grouping them in Bloom's revised matrix (Anderson and Krathwohl, 2014). This matrix contains two dimensions: the knowledge dimension on the vertical axis and the cognitive process dimension on the horizontal axis (Cannon and Feinstein, 2005). This means that the complexity to understand and answer a question increases from left to right, and the complexity of knowledge further increases from top to bottom. The squares in the matrix were left empty by Cannon and Feinstein (2005). We fill each of these squares with typical question keywords, ranging from simple factual questions ("list", "define", "name") to more complex questions ("explain", "analyze", "justify"). These keywords can then in turn be mapped to specific question types. For example, "Who invented..." might be a representative for a factual question in the cognitive process dimension "remember". The three steps to follow are ...

1. filling in the matrix with keywords,
2. assigning categories to question types, and
3. defining the difficulty for the question types.

The question we want to answer is at what level of knowledge and cognitive level multi-hop reasoning is required. The levels could then be used as a basis for classifying responses, as more complex questions will require complex answers. Determining the threshold of complexity in some experiments remains for future work. The approach is also intended to give a very general idea of how to measure question complexity, which is why domain dependence is not considered.

We use existing keywords that we can map to Bloom's Taxonomy and perform classification on a QA dataset. The questions of the dataset are analyzed syntactically such that the model can be independently applied to other domains.

## 4 Proof of Concept

In the following, we will show how to map Bloom's Taxonomy to question difficulty estimation. Therefore, typical question keywords will be filled into the revised matrix of Bloom's Taxonomy. Then, the questions will be tagged with their respective Part-of-Speech (PoS) tags to capture their syntactic features. For classification, a multi-layer perceptron (MLP) will be trained and evaluated on the development data.

### 4.1 Keyword Mapping

To establish a connection between pedagogy and QA systems, we fill in typical question indicator words from educational studies into the revised version of Bloom's Taxonomy. For each slot in the matrix, keywords used by Bloom to estimate question complexity were assigned to their respective categories. This is important because (a) the keywords may appear directly in search queries and (b) the keywords may be used later to assign question words and templates to categories. The results are presented in Table 1.

| The Knowledge Dimension | The Cognitive Process Dimension | | | | |
|---|---|---|---|---|---|
| | 1<br>Remember | 2<br>Understand | 3<br>Apply | 4<br>Analyze | 5<br>Evaluate |
| A<br>Factual | name, list<br>define, label | restate<br>order | state<br>determine | distinguish<br>classify | select<br>according to |
| B<br>Conceptual | identify<br>locate | describe<br>explain | illustrate<br>show | examine<br>analyze | rank<br>compare |
| C<br>Procedural | tell<br>describe | summarize<br>translate | solve<br>demonstrate | deduct<br>diagram | conclude<br>choose |
| D<br>Meta Cognitive | – | interpret<br>paraphrase | find out<br>use | infer<br>examine | justify<br>judge |

Table 1: The knowledge dimension matrix by Cannon and Feinstein (2005) filled with key indicators for complex questions. The complexity ranges from low (top left) to very high (bottom right).

The keywords in the matrix indicate the level of complexity within a search query. While simpler questions are located at the top left, more complex questions are positioned on the bottom right of the table. In the next step, questions from the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) will be extracted, to map the keywords to question terms. This helps to extract a selection of questions and classify it according to the cells in the matrix. The question selection will be described in more detail in the next section.

### 4.2 Question Selection

Next, the keywords from Table 1 are used to extract and classify questions from a QA dataset. The procedure is as follows: For each keyword in the matrix, search through the dataset and extract questions that contain these keywords. For our proof of concept, we searched SQuAD, a reading comprehension dataset consisting of 100,000+ questions from Wikipedia articles. In the dataset, all sentences are searched by keyword and marked correspondingly.

There are two obstacles to overcome. The first is the small amount of about 770 questions that contain keywords. The second is the unequal distribution of samples across the classes. For some categories no questions exist (D5), some classes only have 2 samples (C4) and others are overrepresented (D3) with 414 samples. Two classes have significantly more samples than the rest, namely A1 with 362 samples and D3 with 414 samples. We circumvent both obstacles by transforming the task into a binary classification task and by defining representatives for simple questions (A1) and complex questions (D3). We then argue that the complex samples from D3 will require multi-hop reasoning answers.

To abstract the question structure of the training set, all words are annotated with their respective PoS tag. Since question words may indicate the question complexity, they are included without any adaptations. This will allow us to derive the complexity of a question from its underlying syntactic structure. An example from the class 0 (A1) looks as follows:

| *Name* | an | example | of | a | heavy | isotope |
|---|---|---|---|---|---|---|
| VERB | DET | NOUN | ADP | DET | ADJ | NOUN |

An example from class 1 (D3) shows how question particles remain untagged:

| What | number | is | *used* | in | perpendicular | computing |
|---|---|---|---|---|---|---|
| WHAT | NOUN | AUX | VERB | ADP | ADJ | NOUN |

In the next step, the annotated sentences are used for classifier training. The classifier and the training process are described in the following section.

### 4.3 Question Classification

The question set comprises about 770 samples and is split into 90% training and 10% validation sets. Following the Google developers guide for choosing our classification model, we calculate the samples/number of words per sample. For a ratio smaller than 1,500, they advise to choose a word $n$-gram-based MLP. Therefore, we split the samples into $n$-grams (where $n = \{1, 2, 3, 4\}$) and convert the numbers into vectors. Subsequently, the vectors are scored by importance using tf-idf (short for term frequency-inverse document frequency).

The vectors are fed into the MLP with 3 layers and 64 units and trained for 15 epochs. We added a dropout of 0.2 and early stopping with $patience = 3$ on validation accuracy to prevent overfitting of the model. The results are presented in the next section.

### 4.4 Evaluation

The results show that binary classification, which distinguishes between classes 0 (simple answer) and 1 (multi-hop answer) with A1 and D3 as representatives, yields good results with a simple MLP. The training loss could be reduced after 15 epochs to 0.21 with an accuracy of 0.92, a validation loss of 0.42, and a validation accuracy of 0.85. Figure 1 shows the loss and validation values for each epoch. The best results were obtained with PoS tag $n$-grams where $n = \{1, 2, 3, 4\}$ and a learning rate of 1e-3.

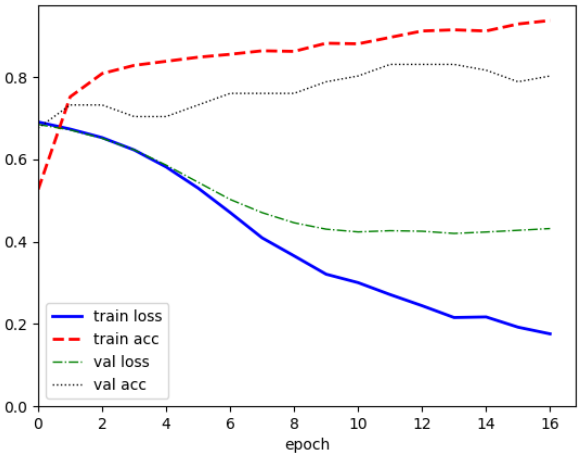


Figure 1: Model loss and accuracy per epoch for 15 epochs after early stopping on validation accuracy with a patience of 3.

When evaluating the model on the validation data, we obtain a weighted F1 value of 0.85 for both classes, with an F1 value of 0.72 for class 0 and 0.88 for class 1. Class 1 achieves the highest recall value of 0.92. A look at the confusion matrix (Figure 2) shows that the vast majority of classes were assigned correctly.

To compute the complexity level, the diagonal of the matrix could be a determinant for the definition of complex questions. For automated calculations, add 1 for each step to the right and down the matrix. If the value is greater than a certain threshold, multi-hop reasoning can be considered.

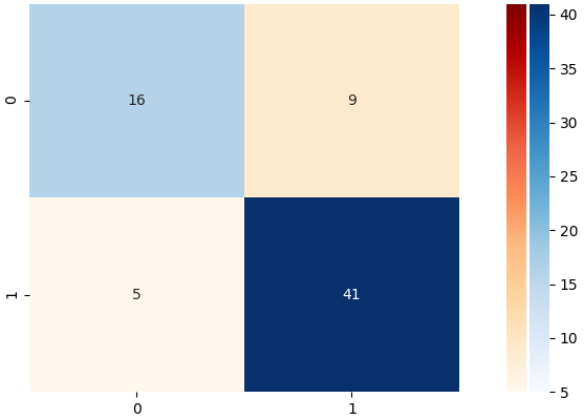


Figure 2: In this confusion matrix for binary classification, 0 represents simple answers (A1 in Bloom's Taxonomy) and 1 means that multi-hop answers are required (D3 in Bloom's Taxonomy).

## 5 Conclusion and Future Work

We have shown that Bloom's revised taxonomy can be transferred from pedagogy to QA systems. The diagonal of the matrix is a determinant for defining complex questions, ranging from simple questions in the upper left to complex questions on the bottom right. For the proof of concept, we added PoS tags to the questions as syntactic information to train a domain-independent classifier for question complexity. We argued that question words also contribute to complexity, so they were not transformed. Although the unequal distribution of the training data only allowed a binary classification for two representative classes A1 and D3, the classifier already provides good results for computing question complexity.

In the future, we plan to collect a larger number of questions from different types of datasets so that a greater diversity of questions is captured. This is crucial for obtaining a diverse data source with a balanced combination of simple and complex questions. It also allows us to expand the question pool so that more classes can be included in the classification. Next to PoS tagging, a wider variety of linguistic features that contributes to the complexity of a question should be considered. This includes the sequence length and the inclusion of semantic information in the classification model. Finally, a user study could help to determine the specific threshold within Bloom's Taxonomy that indicates the need for multi-hop reasoning.

## References


Lorin W. Anderson and David A. Krathwohl, editors. 2014. *A taxonomy for learning, teaching and assessing: A revision of Bloom's*. Pearson Edition.

Amine Bayoudhi, Hatem Ghorbel, and Lamia Hadrich Belguith. 2013. Question Answering System for Dialogues: A New Taxonomy of Opinion Questions. In *International Conference on Flexible Query Answering Systems*, pages 67–78. Springer.

Benjamin S Bloom, Max D Engelhart, EJ Furst, Walker H Hill, and David R Krathwohl. 1956. Handbook I: cognitive domain. *New York: David McKay*.

Hugh M Cannon and Andrew Hale Feinstein. 2005. Bloom beyond Bloom: Using the revised taxonomy to develop experiential learning strategies. In *Developments in Business Simulation and Experiential Learning: Proceedings of the Annual ABSEL conference*, volume 32.

David Eichmann and Padmini Srinivasan. 1999. Filters, webs and answers: The University of Iowa TREC-8 results. In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*. Citeseer.

Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proc. of the 2020 EMNLP*, pages 137–150.

Yang-woo Kim. 2014. Typology of ambiguity on representation of information needs. *Reference and User Services Quarterly*, 53(4):313–325.

Kenneth C Litkowski. 1999. Question-Answering Using Semantic Relation Triples. In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, pages 349–356.

Jing Liu, Quan Wang, Chin-Yew Lin, and Hsiao-Wuen Hon. 2013. Question difficulty estimation in community question answering services. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 85–90.

Sarah KK Luger and Jeff Bowles. 2013. Two methods for measuring question difficulty and discrimination in incomplete crowdsourced data. In *First AAAI Conference on Human Computation and Crowdsourcing*.

Amit Mishra and Sanjay Kumar Jain. 2016. A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3):345–361.

Ulrike Padó. 2017. Question difficulty–how to estimate without norming, how to use for automated grading. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications*, pages 1–10.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392.

Jiankai Sun, Sobhan Moosavi, Rajiv Ramnath, and Srinivasan Parthasarathy. 2018. QDEE: question difficulty and expertise estimation in community question answering sites. In *Twelfth International AAAI Conference on Web and Social Media*.

Håkan Sundblad. 2007. *Question classification in question answering systems*. Ph.D. thesis, Institutionen för datavetenskap.

Jun Suzuki, Hirotoshi Taira, Yutaka Sasaki, and Eisaku Maeda. 2003. Question classification using HDAG kernel. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, pages 61–68.

Toru Takaki. 2000. NTT DATA TREC-9 Question Answering Track Report. In *Proceedings of the 9th Text Retrieval Conference (TREC-9)*.

Deepak Thukral, Adesh Pandey, Rishabh Gupta, Vikram Goyal, and Tanmoy Chakraborty. 2019. DiffQue: Estimating relative difficulty of questions in community question answering services. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(4):1–27.

Quan Wang, Jing Liu, Bin Wang, and Li Guo. 2014. A regularized competition model for question difficulty estimation in community question answering services. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1115–1126.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable NLP. *arXiv preprint arXiv:2102.12060*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380.