Mapping User Intents in Web Search Queries to Types of Commonsense Knowledge

Xiaoao Huang
Supervisor(s): Gaole He, Ujwal Gadiraju
EEMCS, Delft University of Technology, The Netherlands

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Commonsense knowledge is a type of knowledge consisting of facts that humans use every day. Humans make queries in search engines with different user intents, and some of them can be answered by knowledge tuples. Different types of knowledge are stored differently in the knowledge bases. Being aware of the types of commonsense knowledge required to answer the queries can accelerate the process of finding corresponding knowledge for the search engines to give a response to users. For some queries with specific user intents, it is not possible to be answered solely with commonsense knowledge because some analysis and judgment from humans are needed. On the other hand, some queries can be answered with commonsense knowledge tuples and the user intents can have a strong indication of what the knowledge type is required to answer. The research is to look into how to map queries and their user intents to knowledge types and explore the impacts of user intents in the knowledge type classification. There was no existing dataset that had annotations on both user intents and knowledge types. In this work, the described dataset was created. Observations of the created dataset and experiments on three classifiers with accuracy being around 0.99 were conducted. The results show that user intents generally help the classification of the type of commonsense knowledge.

1 Introduction

Users around the world make billions of queries on search engines every single day with different intents [10]. User intents can vary from *Navigational* and *Informational*, to *Transactional* [12], depending on their goal when using the search engine. It is essential for search engines to predict the user intents, in order that what appears in the search results can be optimized by presenting the most relevant content users expect.

To give responses back to users, commonsense knowledge tuples can be used for generating answers to the questions. Commonsense knowledge, which is used and obtained easily by humans, helps humans make sense of everyday situations [6]. FindItOut is a game with purpose (GWAP) to collect commonsense knowledge tuples. It is also the first GWAP to directly collect four types of knowledge including negative knowledge like "Birds cannot fly", and discriminative knowledge, for instance, "The sun is bigger than the moon". Different types of knowledge are stored differently in the knowledge bases. As a result, it takes more time for search engines to look for related knowledge tuple. Knowing what types of commonsense knowledge the queries needed can help make this process faster and reduce the response time. In addition, user intents being extensively studied also set a good basis for this study.

The research question is, how queries and their categorized user intents can be mapped into the knowledge types that FindItOut collected. The hypothesis is that, when including user intents in the training process, the model performs better than the one without help from user intents.

There was no existing dataset that had annotations on both user intents and knowledge types. To answer the research question, a dataset was created and observations on the dataset were made to see how the different knowledge types are distributed in different user intents categories. Experiments are done to figure out whether the accuracy of models improves on knowledge type classification, given their user intents. The accuracy score of using and not using user intents both reached around 0.99 for classifiers. The accuracy of classification with user intents generally outperformed the other when there were fewer samples.

The paper is organized as follows. In Section 2, related work is described. Section 3 presents the methodology used in the research, followed by the discussion about experiments shown in Section 4. Discussion about the results of experiments, implications, and limitations, can be found in Section 5. In Section 6 and Section 7, responsible research and conclusions are presented, respectively.

2 Preliminary Knowledge and Related Work

In this section, preliminary knowledge and related work on natural language processing, the relation of search queries and user goals, question-answering systems, and knowledge bases and knowledge types are presented.

2.1 Natural Language Processing and BERT

In recent years, there has been an exponential growth in the need for accurate text classification, which is a fundamental task in the field of natural language processing [8]. The paper done by Kamran Kowsari et al. has given an overview of different text feature extractions, dimensionality reduction methods, existing algorithms and techniques, and evaluation metrics. One of the most commonly used models to achieve the goals is BERT [4]. BERT stands for Bidirectional Encoder Representations from Transformers, and provides a pre-trained model that can be fine-tuned. BERT is extremely powerful for natural language processing tasks and is chosen because it fits the goal of the research question in which queries and user intents should be analyzed.

2.2 Search Queries and User Goals

This work is related to research conducted by Markus Strohmaier and Mark Kröll, who have looked into how search query is related to common human goals [14]. The researchers have applied an automatic classification approach to search query logs, which largely lowers the costs of knowledge acquisition.

2.3 Question-answering Systems

For the purpose of offering an answer to the user, researchers have created algorithms to map queries to commonsense knowledge in knowledge bases. Relevant research has been done by Abdul Quamar et al. in the question answering system area [11]. With the training of the intents in users' questions, they developed a conversation system that automatically constructs answers to domain-specific questions.

Though there is no question-answering system involved in this work, it is still pertinent to this research with the utilization of user intents in the training process.

2.4 Knowledge Bases and Knowledge Types

ConceptNet [13] is a large-scale commonsense knowledge base to manage textual information and is widely used, and it consists of positive and generative knowledge tuples [14]. There are a number of games with a purpose(GWAP) used for the collection of commonsense knowledge, including Verbosity, RobotTrainer, Virtual Pet, and FindItOut in which two players contribute distinct tuples of knowledge simultaneously [2]. For example, in FindItOut, players, in turn, ask questions and give answers about each other's target card. Their questions and answers can be then collected, organized, and utilized as commonsense knowledge. In the paper about the first GWAP to directly collect discriminative and negative knowledge, FindItOut, discriminative knowledge has been regarded in contrast to generative knowledge. Specifically, generative knowledge has been stated to be about the information about an entity, while discriminative knowledge is for the identification of the difference between entities. Researchers including Hiba Arnaout et al. have also made emphasis on the importance of the existence of negative knowledge in their work [1]. Negativeness in knowledge refers to the invalidity of a tuple to characterize a concept or two compared concepts. This type of knowledge is more likely to bring down the ambiguity and makes it less costly to find answers in the knowledge bases.

As negative and discriminative knowledge types are being brought up by researchers, it takes more time to look for knowledge in the knowledge base to answer the query. To solve the problem, queries can be firstly associated with knowledge types and only certain types of knowledge are searched in the knowledge base. Since with the specific knowledge the query requires being known, it is relatively easy to provide a reasonable answer back to users, exploration of this relation will highly increase the efficiency of providing a response to the query.

Although there has been no research done to directly look into how user intents are associated with the classification of commonsense knowledge types, the text classification techniques, the application of user intents, and knowledge types are still applicable in this paper.

3 Methodology

To build a dataset, queries from MS Marco, Quora, and Ask Reddit were looked through. A subset of the queries was randomly selected from each of the three datasets, and they were combined into a new dataset. In order to obtain a more balanced dataset with a better distribution in both user intents and knowledge types, more queries in specific classes were picked.

Each of the queries in the combined dataset was labeled with its user intent according to the taxonomy created by Jasmine Diaconu [5]. Queries were put into four categories consisting of *Informational*, *Navigational*, *Transactional*, and *Human* which were furthermore divided into more detailed sub-categories.

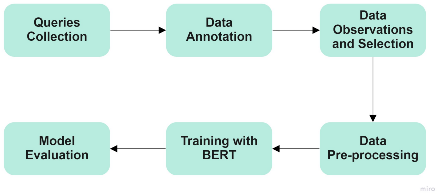

Figure 1: The Process of Experiments

The dataset went through a selection and the samples were put into BERT tokenization and embeddings for the data preprocessing. After the dataset was pre-processed, the BERT pre-trained model was used. The query and the categorized user intent from the query were the features of the model. Aiming to figure out whether the categorized user intents improve the performance of the classification, testing on the model with only the query itself was included. For the purpose of reducing overfitting error and obtaining a better indication of unseen data, 5-fold cross validation was applied. The performance of the machine learning models was measured by the accuracy of the validation sets and test set. The hypothesis was that, with categorized user intents, the performance of the classification enhanced, that is, reaching higher accuracy. A flow chart of how the experiment was done is shown in Figure 1.

4 Experiments and Results

In this section, details of how the experiments were conducted are shown. This includes experimental setup, dataset preparation, observations on the dataset, and the study of the impacts of user intents on knowledge type classification.

4.1 Experimental Setup

In this subsection, how the models were evaluated is described, followed by presenting the hierarchical classifiers and experimental environment.

Evaluation Metrics

In this work, the accuracy value was used for the evaluation of the model performance. Accuracy was calculated by the number of total predictions dividing the number of correct predictions. The accuracy of the validation set and the test set were both taken into consideration. The average score of accuracy was taken for the 5-fold cross validation.

Hierarchical Classifiers

In the initial labeling, there were four types of knowledge, including positive-generative knowledge, negative-generative knowledge, positive-discriminative knowledge, and negative-discriminative knowledge. The problem that arose from this categorization was that the training label was fairly unbalanced. There were much more positive-generative and positive-discriminative knowledge compared to negative-discriminative. In the process of labeling and looking for queries for the creation of the dataset, there was no negative-generative knowledge detected. The unbalance of samples was due to the way and norm users put queries in the web

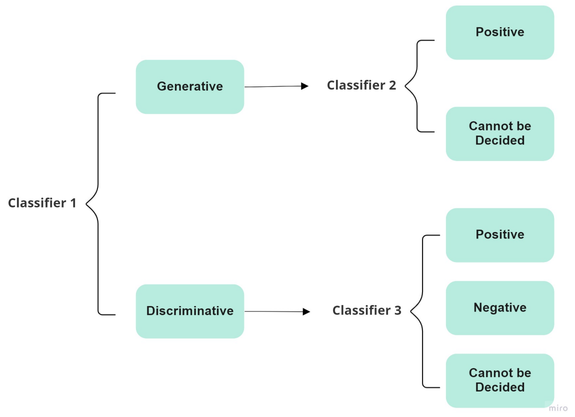

Figure 2: The Hierarchical Classifiers

search engines. It is more likely to search for one entity at a time rather than two, and ask for a positive answer.

To achieve a more balanced knowledge type classification, three classifiers were introduced, as presented in Figure 2. Classifier 1 was used to identify whether the knowledge required to answer the queries was generative or discriminative. After being generative or discriminative was determined, two more classifiers were introduced. These two classifiers decided whether the queries were positive or negative. As mentioned in the previous paragraph, no negative-generative knowledge type was found and thus, the classifier for the generative knowledge had only two labels, that is, positive or cannot be decided. Thus, for classifier 2, it was known that the queries require generative knowledge type, and these queries needed to be identified whether they are positive knowledge types or cannot be decided both with and without user intents. Classifier 3 was for queries that require discriminative knowledge type. It made classification on whether they were positive knowledge type, discriminative knowledge type or cannot be decided.

Experimental Environment

The experiments were done in Google Colab with TPU hardware accelerator. With TPU, the training process was highly speeded up. Libraries used in the experiments includes transformers from Hugging Face, TensorFlow, sklearn, pandas, NumPy, Matplotlib and Keras.

4.2 Dataset preparation

In this subsection, how the dataset was annotated and preprocessed is described.

Dataset annotations

The dataset created was composed of 6460 search queries in total. Among them, 3954 queries are obtained from MS Marco, 1251 from Quora, and the remaining are chosen from Ask Reddit. The reason for choosing a combination of different datasets was to increase the variety in user intents, making sure the size of each category is well-balanced.

The annotation process of the knowledge type was done with the creator of the taxonomy used in this research. The taxonomy was fully studied before the start of annotation. We firstly labeled the dataset alone, followed by a comparison and discussion of the results between the two of us. An agreement from us was reached on the present dataset annotation on user intents.

Data Pre-processing

However, not all of them were suitable for this research. Due to the limitation of the size of *Transactional* and *Navigational* queries, and the fact that they were not questions that required answers, they were ignored in the training process. For *Human* queries, which had adequate quantity in the dataset, led to another issue of being too complicated to answer or the knowledge required to answer them did not belong to commonsense. Thus, only queries classified as *Informational* were considered and labeled with knowledge types.

After filtering, 3916 *Informational* classifiable queries were included in the BERT pre-trained model. In BERT, stop words were automatically handled well, given rather small weights in the model. Additionally, the stop words were able to provide context information, which means that they were as valuable as other non-stop words. Removing stop words from the data or not was not expected to have major benefits for the model performance. Thus, stop words were kept for the experiments. The tokenizer used also took care of lowering the case of the text, so there was no need for data cleaning beforehand.

Furthermore, BERT base model (uncased) was chosen as the tokenizer. The tokenizer was used to split the raw text into smaller units, which were called tokens [9]. The tokenizer took the input text and inserted a special [CLS] token at the beginning of a sentence. It contained no information itself, but as a part of the sentence classification. Similarly, a [SEP] token was appended at the end of each sentence, as a separation of two sentences connected to each other. The sequence length is enforced at a fixed value, by padding the sequence of tokens. In this experiment, the sequence length was set to 20, and it was also the maximum length of text that could be taken into BERT. Any text with a length being less than 20 was filled by [PAD] tokens in the end.

The BERT encoder expected a sequence of tokens. The input text with the length being 20 was passed into three embeddings. Embeddings were vectors that were used to encapsulate words and make the model easier to work with [15]. In token embeddings, each word token was converted into a 768-dimensional vector representation by the token embeddings layer. Segment embeddings represented which sentence the tokens belonged to. Position embeddings were there to present the position of the word within that sentence it belonged to. The results of three embeddings were summed element-wise to produce a single representation. This was the representation that was input to the encoder layer of BERT. The attention mask was included to make sure that the model did not consider the padding tokens. The positions of padding tokens were given 0 in the vector, while other tokens with actual meaning receive 1.

The dataset was split into a training set and a test set. The split ratio was 90% and 10% for the training set and test set respectively. To lower the errors caused by overfitting and obtain a better indication of unseen data, 5-fold cross valida-

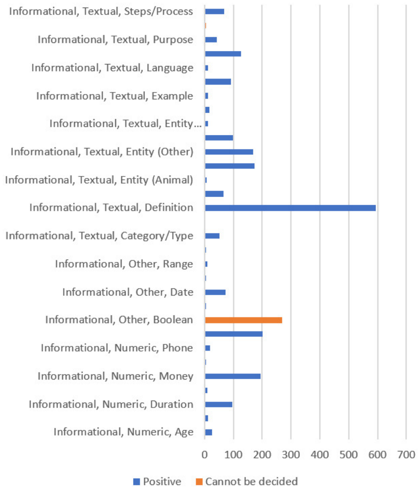

Figure 3: Number of Generative Queries That Are Positive or Cannot Be Decided in Each Category

tion was applied. In order to achieve that, the training set was furthermore divided into 5 groups. In each round, one of the subsets was regarded as the validation set, and the rest of the subsets became the training sets. In the next round, another group of samples was switched to be the validation set and what remained was treated as the training set. This progress was repeated five times before an average of the accuracy in each round could be calculated.

4.3 Dataset Observations

In this subsection, the observations on how different types of knowledge were distributed over user intents categories are shown.

There were 3816 samples used for training the classifier to identify whether the queries need generative or discriminative knowledge tuples. Among 1439 queries that were labeled as discriminative, 1427 of them had *Comparison* as the user intents. That took up 99% of the discriminative samples, which gave the model a tremendously strong indication of what queries should be labeled as discriminative. The reason for the large portion of discriminative queries being in the *Comparison* category was that the characteristic of the discriminative knowledge tuples always involved two entities.

The next step was to identify if a generative query in the 2477 samples was positive, or it was unclear to classify. It also resulted in an unbalanced distribution. 270 out of 272 generative queries had their user intents as *Boolean*. It was apparently the case because whether the question required a positive or negative knowledge tuple could only be decided when it got access to the actual knowledge bases. Having *Selection* user intents could lead to the knowledge required being difficult to decide as well. It can be seen from Figure 3 that except for the queries with *Boolean* or *Selection* user intents, other queries in the dataset were very likely to be positive.

The last classifier identified whether a sample labeled as discriminative was positive, negative, or cannot be decided. It only dealt with 1439 samples and five categories of user intents. The samples for this classifier were more balanced than the previous two. When taking comparison, which had the biggest portion in discriminative queries, as an example, it can be seen from Figure 4 that there were probability of all three labels in this category. It could be reasonable if user intents in this classifier had less significant impact on the classification. More specific tables for the distribution of the knowledge types in different user intents can be found in Appendix A.

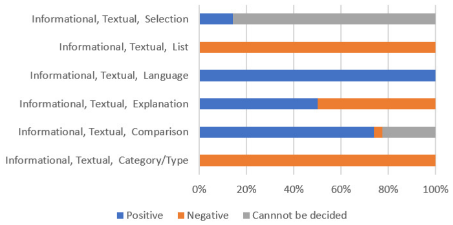

Figure 4: Portions of Positive and Negative Discriminative Queries in Each Category

4.4 Impacts of User Intents on Knowledge Type Classification

In this subsection, experiments of three classifiers on knowledge type classification with and without user intents are presented.

Impacts of User Intents on The Classifier That Identifies between Generative and Discriminative Knowledge

In total, 3916 samples went through the model BERT provided. Both using and not using user intents as part of the features resulted in an accuracy of over 0.99. The reason for the extreme high accuracy was because of the strong implication of *Comparison* user intents.

To better look into the impact of user intents in the classification, a subset of the 3916 samples were chosen. The subset was made up of 20 samples in each user intents category, that is, 630 samples in total.

There were 109,483,778 parameters trainable in BERT model. Making the parameters in the first layer non-trainable decreased the accuracy by more than 30%, while dropping out the 1538 parameters in the third layer did not noticeably affect the model performance overall. To make the best use of the BERT model, all the parameters were kept trainable. The hyperparameters to be fine-tuned included the number of epochs and batch size in the model. The epoch represents the number of times the learning algorithm should pass through the entire training dataset, while the batch size was the number of a set of training samples to work through before an update of the internal parameters [3]. The number of epochs was set to 1, which was chosen to avoid overfitting issues. The batch size should be a power of 2 to make full use of the GPUs processing and was tuned to be 32 as recommended [7].

Adam optimizer was selected to minimize the loss and reach the optimal point in the loss function.There are also few hyperparameters in the Adam optimizer, including learning rate, decay rate, and momentum. One of the hyperparameters that affect the results most was the learning rate. It was set to 3e-5.

Table 1 shows that with user intents, the accuracy results outperformed the model without considering user intents in 5-fold cross validation. Furthermore, the performance of the model on test set for having user intents was better than not having user intents, with accuracy being 0.9677 and 0.9516 respectively. It was indicated that for this classifier, user intents played an important role in helping the knowledge type classification.

Table 1: Accuracy Scores of Models with User Intents and without User Intents of A Smaller Dataset for Classification of Generative or Discriminative Knowledge Types

Datasets	With User Intents	Without User Intents
Fold 1	0.9732	0.9732
Fold 2	0.9643	0.9640
Fold 3	0.9732	0.9554
Fold 4	0.9732	0.9550
Fold 5	1.000	0.9820
Average	0.9768	0.9659
Test Set	0.9677	0.9516

Impacts of User Intents on The Classifier That Identifies Whether the Generative Knowledge Is Positive or Cannot Be Decided

The experiments of this classifier were conducted with the same parameters. The accuracy results on the full dataset turned out to be as high as the previous classifier, with 0.9960 for both including user intents as inputs or not. In order to explore the associations between user intents in these queries and knowledge types, a similar approach as the experiments on the previous classifier was done for this one. There were altogether 630 samples evenly and randomly selected from 31 user intent categories, forming a smaller group for training. The model with user intents, again, outperformed the one without the help of user intents, with an accuracy being 0.9839 and 0.9695 respectively as the results of 5-fold cross validation. Both models achieved over 0.99 accuracy for the test set, which meant that, the models were pretty well-trained.

Impacts of User Intents on The Classifier That Identifies Whether the Discriminative Knowledge Is Positive, Negative or Cannot Be Decided

For samples used in experiments on this classifier, there were only six user intents involved, including *Category/Type*, *Comparison*, *Language*, *List*, and *Selection*. The initial experiment's result on this classifier differed from the previous two and reached a relatively lower accuracy score. To rise the training model accuracy, the parameters were adjusted. The batch size was set to 8 while the learning rate and the number of epochs were retained. This tuning increased the accuracy from 0.9838 to 0.9907 for samples with user intents, while the accuracy of the model without using user intents grew from 0.9807 to 0.9946. When a smaller sample group was applied for the experiment, the accuracy for the test set was no longer stable, while the performance of the validation set was as high as the original dataset. It showed that the overfitting problem occurred as the models fit exactly against the training data and could not predict unseen data well enough. The model without user intents, furthermore, did not always underperform the other one. On the contrary, it outperformed the model, which took the help of user intents with more than 0.1 on accuracy.

5 Discussion and Further Work

The accuracy results of the three classifiers indicate that the models are all capable of predicting the knowledge types of the queries that need to be answered. With the most accuracy scores above 0.99, the training samples are adequate and have clear patterns for the models to learn. Having a smaller dataset resulted in a worse performance, which was expected.

The results of the observation of the dataset and the experiments show that user intents have associations with knowledge types. More specifically, being aware of the user intents helps the model classify whether the queries require generative or discriminative knowledge, and assists in the identification of whether the generative knowledge is positive. However, regarding the classification of whether the discriminative knowledge is positive, negative, or cannot be decided, user intents do not play a key role in improving the model's performance.

With the results of this work, when search engines use knowledge bases that include negative and discriminative tuples, they could firstly apply classification on user intents, which will help with the classification of the knowledge type. In this way, the response of these search engines will speed up when the knowledge types of the queries are known.

Despite the implications of the research, there are still some limitations. Firstly, the dataset was not well-balanced in respect of knowledge types. There were not many samples categorized as negative discriminative compared to others. Moreover, certain types of samples lacked diversity, especially the discriminative ones. The queries categorized as discriminative usually follow one of the few patterns of phrasing. This result was due to the definition of discriminative knowledge being about a comparison of two entities. Thus, limited words like "different", "similar", and "than" usually imply that discriminative knowledge is needed. In the future, the variance of discriminative knowledge should be looked into, and more diverse samples can be taken into the training. Besides, in the fine-tuning process, only a few distinctive combinations of hyperparameters and weights were compared, and the parameters were set with fairly decent performance. In future work, more combinations can be tried out to achieve more accurate classification.

6 Responsible Research

Concerning ethics, the dataset was created using anonymous search queries selected from public datasets including MS Marco, Quora, and Ask Reddit. No queries in the dataset of this research came from outside the mentioned datasets. Furthermore, user queries used in the research do not collect, user, or store any sensitive or personal information.

Additionally, the conducted experiments are reproducible and repeatable. The environmental setup of the code related in this work was introduced in Section 4.1. This includes the platform the code should be running and the libraries used. The hyperparameters set for the three classifiers are described in Section 4.4. The code and dataset involved in the classification were uploaded to GitHub in a public repository.

7 Conclusions

In this research, a dataset containing web search queries, their user intents, and knowledge types they required to be answered were annotated and used. It can be seen that the queries labeled as *Comparison* have an extremely large probability of requiring discriminative knowledge, while most of the queries that look for generative knowledge but cannot decide whether positive or negative knowledge is needed fall into the *Boolean* user intents. A model from BERT was used for the knowledge type classification and achieved an accuracy of 0.99 after fine-tuning. Experiments on the training models for the classification with and without user intents confirm that there are connections between the existence of user intents and the classification performance. With user intents, the accuracy of classification is generally higher than not having them.

References

[1] Hiba Arnaout, Simon Razniewski, Gerhard Weikum, and Jeff Z Pan. Wikinegata: a knowledge base with interesting negative statements. *Proceedings of the VLDB Endowment*, 14(12):2807–2810, 2021.

[2] Agathe Balayn, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. Ready player one! eliciting diverse knowledge using a configurable game. In *Proceedings of the ACM Web Conference 2022*, pages 1709–1719, 2022.

[3] Jason Brownlee. Difference between a batch and an epoch in a neural network. Available at https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/ (Acessed: 07-06-2022), 2018.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Jasmine Diaconu. A comprehensive taxonomy of user intents for search queries. Manuscript in preparation, 2022.

[6] Filip Ilievski, Alessandro Oltramari, Kaixin Ma, Bin Zhang, Deborah L McGuinness, and Pedro Szekely. Dimensions of commonsense knowledge. *Knowledge-Based Systems*, 229:107347, 2021.

[7] Ibrahem Kandel and Mauro Castelli. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express*, 6(4):312–315, 2020.

[8] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

[9] Aravindpai Pai. What is tokenization in nlp? here's all you need to know. Available at https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/(Accessed: 06-06-2022), 2021.

[10] Meg Prater. 25 google search statistics to bookmark asap. Available at https://blog.hubspot.com/marketing/google-search-statistics (Accessed: 30-05-2022), 2021.

[11] Abdul Quamar, Chuan Lei, Dorian Miller, Fatma Ozcan, Jeffrey Kreulen, Robert J Moore, and Vasilis Efthymiou. An ontology-based conversation system for knowledge bases. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 361–376, 2020.

[12] Jeremy Smith. The conversion optimization guide to user intent. Available at https://www.crazyegg.com/blog/guide-user-intent/ (Acessed: 30-05-2022), 2014.

[13] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[14] Markus Strohmaier and Mark Kröll. Studying databases of intentions: do search query logs capture knowledge about common human goals? In *Proceedings of the fifth international conference on Knowledge capture*, pages 89–96, 2009.

[15] Adith Narein T. Embeddings in bert. Available at https://iq.opengenus.org/embeddings-in-bert/(Accessed: 06-06-2022).

A Distribution of Knowledge Types in Different User Intents

Table 2 shows the number of generative queries and discriminative queries in each category of user intents. Table 3 illustrates the number of discriminative queries that are positive, negative or cannot be decided in each category. Table 4 presents the number of generative queries that are positive or cannot be decided in each user intents category.

Table 2: The Number of Generative and Discriminative Queries in Each User Intents Category

User Intents	Generative	Discriminative
Informational, Numeric, Age	25	0
Informational, Numeric, Conversion	11	0
Informational, Numeric, Duration	96	0
Informational, Numeric, Frequency	10	0
Informational, Numeric, Money	194	0
Informational, Numeric, Percentage	4	0
Informational, Numeric, Phone	20	0
Informational, Numeric, Quantity	202	0
Informational, Other, Boolean	270	0
Informational, Other, Code	6	0
Informational, Other, Date	72	0
Informational, Other, Formula	4	0
Informational, Other, Range	10	0
Informational, Other, Time	4	0
Informational, Textual, Category/Type	52	1
Informational, Textual, Comparison	1	1427
Informational, Textual, Definition	595	0
Informational, Textual, Description	66	0
Informational, Textual, Entity (Animal)	8	0
Informational, Textual, Entity (Location)	174	0
Informational, Textual, Entity (Other)	169	0
Informational, Textual, Entity (Person)	98	0
Informational, Textual, Entity (Temporal)	13	0
Informational, Textual, Entity (Weather)	17	0
Informational, Textual, Example	12	0
Informational, Textual, Explanation	92	2
Informational, Textual, Language	12	1
Informational, Textual, List	126	1
Informational, Textual, Purpose	43	0
Informational, Textual, Selection	4	7
Informational, Textual, Steps/Process	67	0
Total	2477	1439

Table 3: The Number of Discriminative Queries That Are Positive, Negative or Cannot Be Decided in Each User Intents Category

	Positive	Negative	Cannot Be Decided
Informational, Textual, Category/Type	0	1	0
Informational, Textual, Comparison	1053	53	321
Informational, Textual, Explanation	1	1	0
Informational, Textual, Language	1	0	0
Informational, Textual, List	0	1	0
Informational, Textual, Selection	1	0	6
Total	1056	56	327

Table 4: The Number of Generative Queries That Are Positive or Cannot Be Decided in Each User Intents Category

User Intents	Positive	Cannot Be Decided
Informational, Numeric, Age	25	0
Informational, Numeric, Conversion	11	0
Informational, Numeric, Duration	96	0
Informational, Numeric, Frequency	10	0
Informational, Numeric, Money	194	0
Informational, Numeric, Percentage	4	0
Informational, Numeric, Phone	20	0
Informational, Numeric, Quantity	202	0
Informational, Other, Boolean	0	270
Informational, Other, Code	6	0
Informational, Other, Date	72	0
Informational, Other, Formula	4	0
Informational, Other, Range	10	0
Informational, Other, Time	4	0
Informational, Textual, Category/Type	52	0
Informational, Textual, Comparison	1	0
Informational, Textual, Definition	595	0
Informational, Textual, Description	66	0
Informational, Textual, Entity (Animal)	8	0
Informational, Textual, Entity (Location)	174	0
Informational, Textual, Entity (Other)	169	0
Informational, Textual, Entity (Person)	98	0
Informational, Textual, Entity (Temporal)	13	0
Informational, Textual, Entity (Weather)	17	0
Informational, Textual, Example	12	0
Informational, Textual, Explanation	92	0
Informational, Textual, Language	12	0
Informational, Textual, List	126	0
Informational, Textual, Purpose	43	0
Informational, Textual, Selection	2	2
Informational, Textual, Steps/Process	67	0
Total	2205	272