A comprehensive taxonomy of user intents for search queries

CSE3000 - Research Project

Jasmine Diaconu

Supervisor: Gaole He
Responsible professor: Ujwal Gadiraju

TU Delft

1. Introduction

Search engines endeavor to propose results based on the **user intents** behind search queries.

User intent **categorization** is fundamental to achieving the goal.

Attempts on categorizing user intents date back to the early 2000s, with Broder's **taxonomy** [1].

Drawback: user intents change over time and taxonomies need to adapt to those changes.

2. Research question

How to categorize queries into user intents?

3. Method

DATA COLLECTION

5,000 queries: 50% from **MS Marco** [2], 25% from **Quora** [3] and 25% from **AskReddit** [4].

TAXONOMY COMPOSITION

4-layer hierarchical taxonomy to categorize the collected queries.

DATA LABELING & PRE-PROCESSING

Query labeling: label the dataset according to the categories in the taxonomy.
Text pre-processing: lower-case, punctuation removal, tokenization, and vectorization.

DATASET PARTITIONING

Three splitting strategies:

1. **Full dataset:** train set 70% - validation set 10% - test set 20%
2. **Active Learning with Uncertainty Sampling:** train set 10% - validation set 10% - test set 20%
3. **Active Learning with Random Sampling:** train set 10% - validation set 10% - test set 20%

CLASSIFICATION & EVALUATION

Three Deep Learning models for classification and evaluation:

1. **MLP** (Multilayer Perceptron)
2. **LSTM** (Long-Short Term Memory)
3. **BERT** (Bidirectional Encoder Representations from Transformers)

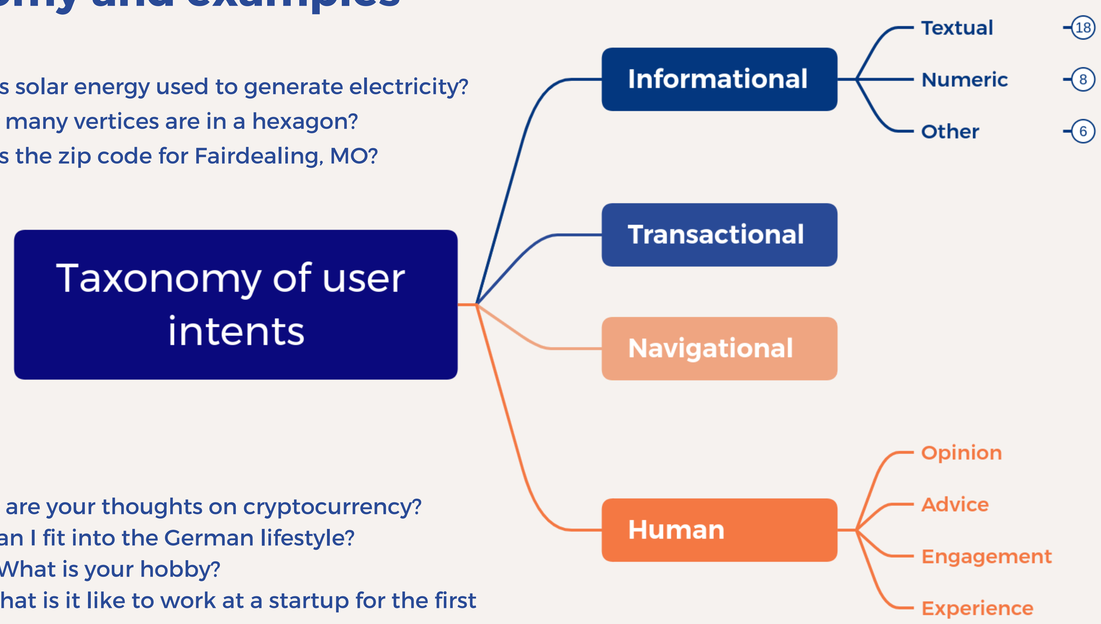
References

[1] Andrei Broder. A taxonomy of web search. SIGIR Forum, September 2002
[2] MSMarco dataset. 2021. Available at https://public.ukp.informatik.tu-darmstadt.de/thakur/BEIR/datasets/
[3] Quora dataset. 2021. Available at https://public.ukp.informatik.tu-darmstadt.de/thakur/BEIR/datasets/
[4] AskReddit dataset. 2021. Available at https://huggingface.co/datasets/SocialGrep/one-million-reddit-questions

4. Taxonomy and examples

Informational

- **Textual:** How is solar energy used to generate electricity?
- **Numeric:** How many vertices are in a hexagon?
- **Other:** What is the zip code for Fairdealing, MO?



Human

- **Opinion:** What are your thoughts on cryptocurrency?
- **Advice:** How can I fit into the German lifestyle?
- **Engagement:** What is your hobby?
- **Experience:** What is it like to work at a startup for the first time?

5. Results

Model	Strategy	Loss	Accuracy	False Negatives	False Positives
MLP	Active Learning with Uncertainty Sampling	0.51	0.73	269	0
LSTM		0.33	0.90	3	96
BERT		**0.12**	**0.97**	**33**	**36**
MLP	Active Learning with Random Sampling	0.33	0.89	46	65
LSTM		0.32	0.88	82	43
BERT		**0.18**	**0.96**	**31**	**42**
MLP	Full dataset	0.34	0.89	38	74
LSTM		0.25	0.92	51	29
BERT		**0.04**	**0.99**	**9**	**10**

Binary classification of Informational and Human queries

6. Conclusion

- The taxonomy is effective at distinguishing Informational and Human queries
- BERT is the classifier with the best performance overall
- Active Learning can achieve good results with less data labeling

7. Future work

- Add Transactional and Navigational queries to the dataset
- Balance the dataset over the layers
- Explore other strategies such as **In-Context Learning** and **Few-Shots Learning**