ORCAS-I query intent predictor as component of TIRA

Daria Alexander[1], Wojciech Kusa[2] and Arjen P. de Vries[1]

[1]Radboud University, Houtlaan 4, 6525 XZ Nijmegen, Netherlands
[2]TU Wien, 1040 Vienna, Favoritenstraße 9-11

Abstract

We present a query intent predictor that is based on Snorkel weak supervision approach. After using it on ORCAS dataset and conducting a series of experiments with a variety of machine learning models we found that the results produced by Snorkel were not outperformed by these competing approaches and can be considered state-of-the-art. The advantage of a rule-based approach like Snorkel's is its efficient deployment in an actual system, where intent classification would be executed for every query issued. When used as a component of TIRA/TIREX platform, our query intent predictor was shown to be applicable to other IR benchmark datasets. Also, we found out that the awareness of the intent overall can improve ranking results for informational intent and its subcategory factual intent.

Keywords

snorkel, weak supervision, intent labelling, web search

1. Introduction

When a user is typing a search query there is usually a specific intent behind it: to explore a topic, to get an answer to a question or to find a particular site. Understanding the intent is beneficial for the user because it can provide better results according to the search goal.

While manual classification provides more accurate labels, it can be very challenging. A weak supervision approach allows avoiding hand labelling of large datasets. To make the labelling easier, we created a query intent predictor that uses Snorkel [1]. We annotated ORCAS dataset [2] which has 18 million connections to 10 million distinct queries and released the ORCAS-I dataset [3] that provides intent categories along with the queries. We also trained 5 machine learning models on the 2-million items subset of ORCAS dataset. The results showed that benchmark models do not significantly outperform the Snorkel classifier, because the lack of external knowledge and the size of the queries (mean length 3.25 words).

To allow the simple re-use of the query intent in diverse retrieval scenarios, we submitted our query intent predictor as software submission to the Workshop on Open Web Search [4]. We dockerized the predictor, submitted it to TIRA [5]/TIREx [6] and run it on all datasets included in TIREx. The experiments showed that our query intent predictor is applicable to other IR benchmark datasets. However, compared to ORCAS-I dataset, the queries whose goal is to find a specific site (navigational) are underrepresented in 40 datasets available in TIRA. The number of navigational queries rises when the dataset has URLs, like TREC Web test collections.

The ranking experiments with TIRA/TIREX showed that, despite the limited information that a short query can provide, the models that understand the semantics of the query are preforming better than the baseline if the queries have informational intent, especially in cases when the user is seeking some specific facts or pieces of information.

WOWS 2024: 1st International Workshop on Open Web Search at ECIR 2024, March 28, 2024

daria.alexander.ru.nl (D. Alexander); wojciech.kusa@tuwien.ac.at (W. Kusa); arjen.devries@ru.nl (A. P. d. Vries)

2. Methods overview

2.1. Taxonomy

To classify the queries in ORCAS dataset according to user intent we created a taxonomy based on Broder's [7] taxonomy that divides intent into three levels: informational, navigational and transactional. Moreover, we added two subcategories in the informational class, factual and instrumental. We provide the definitions of the categories.

- **Navigational intent**: the immediate intent is to reach a particular website [7];
- **Transactional intent**: locate a website with the goal to obtain some other product, which may require executing some Web service on that website [8];
- **Informational intent**: locate content concerning a particular topic in order to address an information need of the searcher [8];
  - **Factual intent**: locate specific facts or pieces of information [9];
  - **Instrumental intent**: the aim is to find out what to do or how to do something [10];
  - **Abstain**: everything inside the informational category that is not classified as factual or instrumental.

To create labelling rules for Snorkel we used the characteristics suggested by Jansen et al. [8]. They utilise established heuristics and keywords to construct rules for the prediction of the intent category for informational, navigational and transactional intent. We re-evaluated those characteristics and added some new ones. For example, we classified the queries that start with question words "what is", "what are", "do" and "does" as factual, because those questions usually need specific pieces of information as answers. For the same reason, we considered the queries that contain words "definition" and "meaning" as belonging to factual intent. The queries that started with a verb were classified as instrumental: (e.g. "bake a pork chop").

Also, we used Levenshtein similarity ratio to determine if the query is navigational. We wanted to understand whether the domain part of the URL and the query are similar. The Levenshtein similarity ratio is computed according to the following formula:

$$\frac{|a| + |b| - \mathrm{Lev}(a, b)}{|a| + |b|}$$

Here, $\mathrm{Lev}(a, b)$ is Levenshtein distance (the minimum number of edits required change a one-word sequence into the other) and $|a|$ and $|b|$ are lengths of sequence *a* and sequence *b* respectively. A threshold on Levenshtein ratio was empirically established at 0.55, which means that if the query and the domain name were 55% or more similar they were classified as navigational.

According to Jansen et al., the queries that do not meet criteria for navigational or transactional have informational intent. Thus, we decided to make abstain subcategory part of the informational category. However, we could not establish consistent automatic characteristics for this group of queries because we could not find any reliable patterns in them.

More details about the intent labelling characteristics for ORCAS-I query intent predictor are provided in [3].

2.2. ORCAS and ORCAS-I datasets

The ORCAS dataset is part of the MS MARCO datasets (Microsoft) and is intended for non-commercial research purposes. It contains 18.8 million clicked query-URL pairs and 10.4 million distinct queries. The dataset has the following information: *query ID*, *query*, *document ID* and *clicked URL*. The documents that the URLs lead to come from TREC Deep Learning Track.

This dataset was aggregated based on a subsample of Bing's 26-month logs to January 2020. The creators of the dataset applied several filters to the log. Firstly, they only kept query-URL pairs where the URL is present in the 3.2 million document TREC Deep Learning corpus. Secondly, they applied a *k*-anonymity filter and only kept queries that are used by *k* different users. Finally, offensive queries such as hatred and pornography were removed.

ORCAS-I dataset is the version of ORCAS dataset annotated with user intent that we released [11]. For Snorkel labelling we used a two-million sample of the ORCAS dataset (ORCAS-2M). In order to evaluate the performance of our weak supervision approach, we manually created a test set collection. We randomly selected 1000 queries from the original ORCAS dataset that were not in the ORCAS-I-2M dataset. The test set was annotated by two IR specialists using the open-source tool *Doccano*[1]. For inter-annotator agreement on the test set, the Cohen Kappa statistic was 0.82. We call this manually annotated dataset ORCAS-I-gold.

2.3. Creating Snorkel labelling functions

In machine learning terms, our intent taxonomy could be represented as a two-level, multi-class classification problem. Snorkel has originally only been implemented to handle annotations for single-level classification problems. As our taxonomy is hierarchical, we needed to define two layers of Snorkel labelling functions.

We defined the first level of labelling functions to distinguish between navigational and transactional intents. All the queries that could not fit into one of these two categories were classified as informational intent in our taxonomy. Based on our user intent classification characteristics, we created four labelling functions for navigational queries and three functions for transactional queries.

On the second level, we defined labelling functions to cover factual and instrumental intents. Similar to the previous step, we designed nine factual and four instrumental labelling functions. All queries that were not assigned a label from the two layers of Snorkel got an abstain category.

We initially used *Spacy*'s *en_core_web_lg* language model to identify part of speech information and to detect named entities. After initial analysis, this proved to generate too many false negatives, especially for the detection of verbs. For example, the queries “change display to two monitors” and “export itunes library” were misclassified as abstain, because the verbs “change” and “export” were labelled as nouns. In the final version, we decided to use a list of the 850+ common verbs with which we obtained comparable coverage with fewer false positives. Eventually, we only have used *Spacy* for a labelling function where queries begin with the “-ing” form of the verb.

2.4. Training Snorkel

To obtain a final prediction score, we run independently two levels of Snorkel annotations. Based on our classification that all non-transactional, non-navigational queries are informational, for the second level prediction, we use all the queries which were assigned abstain from the first level. In order to conduct label aggregation, we experiment with both the LabelModel and MajorityLabelVoter methods implemented in Snorkel. LabelModel estimates rule weights using an unsupervised agreement-based objective. MajorityLabelVoter creates labels by aggregating the predictions from multiple weak rules via majority voting. We test their predictions on the test dataset using default hyperparameters. Results are presented in Table 1.

Table 1

Comparison of Snorkel labelling models results on ORCAS-I-gold

Model	Metric	Precision	Recall	F1score
Majority Label Voter	Macro avg	.780	.763	.771
	Weighted avg	.786	.783	.783
Label Model	Macro avg	.737	.773	.750
	Weighted avg	.779	.770	.772

[1] https://github.com/doccano/doccano

After analysis of results on the testset, MajorityLabelVoter achieved higher scores for all measures except macro average recall. Therefore, we decided to use it to obtain the final labels for the ORCAS-I-2M dataset. MajorityLabelVoter has the additional benefit that it provides more explainable results, as for every query, the user can be presented with the raw aggregation of the single labelling functions.

2.5. Results

2.5.1. Label distribution

Table 2 shows the label distribution in ORCAS-I dataset compared to the 40 test collections present in TIRA[2]. As for ORCAS dataset, the only underrepresented category compared to gold is the navigational intent, which contains 1.6% less items than in the manual labelling approach. The label distribution from Snorkel for ORCAS-2M and ORCAS-18M is comparable, so the 2M sample chosen from the full ORCAS dataset is representative.

As for the test collections present in TIRA, the categories that are comparable to ORCAS-I label distribution are instrumental and transactional. TIRA datasets contain much less navigational queries. Unlike ORCAS, the majority of those datasets do not have URLs, which can explain far fewer navigational queries. As for factual queries, ORCAS-I intent predictor often relies on the form of the query (e.g. starts with "what is", "who is"), which means that some datasets, like ClueWeb collections or BEIR Nutrition Facts Corpus would have fewer queries that belong to this intent.

Table 2
Label distribution in ORCAS-I dataset and TIRA datasets

Label distribution	ORCAS-I gold	ORCAS-I-2M	ORCAS-I-18M	Datasets in TIRA
Navigational	17.10%	14.48%	14.51%	1.83%
Transactional	4.30%	4.17%	4.16%	4.49%
Informational (all)	78.60%	81.35%	81.33%	93.46%
-Instrumental	5.90%	5.82%	5.81%	5.16%
-Factual	36.30%	35.35%	35.35%	17.20%
-Abstain	36.40%	40.18%	40.17%	71.31%

Having a URL is crucial for the navigational intent, that is why better predictions are made in the post-retrieval stage. Comparing the label distribution in TREC Web test collections (Table 3) with and without URL shows us that no navigational intent is detected in the pre-retrieval stage. By contrast, it is detected for all three collections in the post-retrieval stage. It shows that all the elements that could help to disambiguate the intent of the query are important.

Table 3
Label distribution in TREC Web test collections (2011-2014)

Label distribution	TREC Web11		TREC Web12		TREC Web13		TREC Web14	
URL	yes	no	yes	no	yes	no	yes	no
Navigational	18%	-	24%	-	8%	-	4%	-
Transactional	-	-	-	-	-	-	-	-
Informational (all)	82%	100%	76%	100%	92%	100%	96%	100%
-Instrumental	6%	6%	6%	8%	2%	2%	10%	10%
-Factual	8%	4%	10%	10%	18%	18%	6%	6%
-Abstain	68%	90%	60%	82%	72%	80%	80%	84%

[2] Vaswani dataset; Cranfield dataset; TREC Genomics (2004, 2005); TREC Disks 4 and 5; TREC Terabyte (2004-2006); TREC Web(2004-2006,2009-2013); BEIR Nutrition Facts Corpus; TREC Precision Medicine track (2017,2018); TREC Common Core 2018; TREC COVID (CORD19); TREC Medical Misinformation 2019; TREC Deep Learning passage (MS MARCO) (2019,2020); Touche (2020,2021); ANTIQUE; TREC Tip-of-the-tongue 2023; CLEF LongEval 2023.

2.5.2. Comparing Snorkel to benchmark models

We benchmark five different models by training them on ORCAS-I-2M: Logistic regression, Support Vector Machine, fastText, BERT and xtremedistil (more details about the hyperparameters can be found in [3]). To train those models we use two types of training data: just the query and query plus URL. URL features help to improve classification effectiveness. Table 4 shows that when we eliminate URL features from Snorkel (we mute or change the labelling functions that are using URLs) especially recall is reduced. Same as for TREC Web test collections, this is particularly noticeable for the navigational category, for which recall drops from 0.73 to 0.35.

Table 4
Macro average scores comparison for all benchmark models trained on intent categories. Underlined scores indicate the highest score within the different input features for each model, bold values indicate the highest score overall.

Model	Input features	Precision	Recall	F1-score
Snorkel	query	.771	.648	.667
	query + URL	.779	764	.770
Logistic regression	query	.701	.611	.643
	query + URL	.714	.689	.700
SVM	query	.735	.689	.703
	query + URL	.782	.759	.767
fastText	query	.694	.643	.660
	query + URL	.768	.753	.758
BERT	query	.742	.705	.717
	query + URL	**.789**	.764	**.774**
xtremedistil	query	.725	.691	.696
	query + URL	.781	**.765**	.772

We hypothesise that as we take URL features into account for the Snorkel classifier, models that train on queries and URLs will outperform the models that train on queries only. This hypothesis is confirmed for all the models, especially for fastText and xtremdistil. Also, SVM, BERT and xtremdistil show improvements on recall for query-only when compared to Snorkel. It indicates that the models learn well from the labels assigned by the Snorkel query and URL functions, even if they are trained only on queries.

None of our benchmark models significantly outperforms our Snorkel baseline when trained on queries and URLs. This could have been an expected behaviour when comparing two models, one being the teacher and the other the student who learned only from this one teacher, without any external knowledge. We also hypothesise that transformer-based models cannot express their full power because the input sequences are, on average, very short and lack context.

2.5.3. Ranking experiments

For ranking experiments we decided to focus on the informational intent in the taxonomy, because navigational and transactional intent are underrepresented in the datasets availiable in TIRA. Some collections do not have navigational and transactional queries at all (TREC Genomics 2004 and 2005, TREC 2018 Core track) some only have a few navigational and transactional queries (CLEF Longeval 2023). For the experiments we chose the TREC COVID collection and TREC Deep Leairning track 2019 and 2020 passage collections because they provided suficcient samples of instrumental, factual and abstain queries for the evaluation of the models' performance.

We tested the BM25 baseline against two ranking models. ColBERT [12] introduces a late interaction architecture that independently encodes the query and the document using BERT and then employs an interaction step that models their fine grained similarity. MonoT5 [13] is a sequence-to-sequence model that uses a similar masked language modeling objective as BERT to pretrain its encoder–decoder architecture.

Table 5
Ranking according to user intent in different test collections (metric: ndcg@10). Statistical significance is shown with * for the best results compared to BM25. It was measured with a Tukey test ($p < 0.05$).

	Informational(all)	Instrumental	Factual	Abstain
TREC COVID				
BM25	0.281	0.094	0.223	0.306
ColBERT	0.574*	0.184	0.753*	0.548*
MonoT5	**0.688***	**0.256**	**0.846***	**0.669***
TREC 2019 Deep Learning Track passage				
BM25	0.480	**0.897**	0.460	0.519
ColBERT	**0.695***	0.780	0.708*	0.622
MonoT5	0.694*	0.803	**0.723***	**0.650**
TREC 2020 Deep Learning Track passage				
BM25	0.505	0.525	0.508	0.471
ColBERT	**0.689***	0.504	0.697*	0.642
MonoT5	0.695*	**0.698**	**0.701***	**0.636**

Table 5 shows that Monot5 and ColBERT outperform the BM25 baseline for informational queries in general. It gives better results for abstain and factual categories in TREC Covid collection and for factual category in TREC Deep Learning Track passage collections. The findings for informational queries in general and for factual queries are confirmed for TREC Deep Learing Track passage collections. It indicates that, despite the short size of the queries and the lack of context, attention-based rankers are performing better for this type of queries. As most of the factual queries are formulated as questions (e.g. "what is the most popular food in switzerland?"), those models still get benefits from the semantics and the structure of the queries.

3. Conclusion

After creating the ORCAS-I query intent predictor, we used it for TIRA test collections. In terms of the label distribution, TIRA datasets have fewer navigational queries than ORCAS-I dataset, as many of them do not have a URL field. Comparing the pre-retrieval with the post-retrieval intent for TREC Web collections showed that having a query and URL for navigational category provides better results than having only a query, which confirms our previous findings that Snorkel and benchmark models trained on it perform better when having a URL. As for ranking experiments, attention-based rankers performed better than the BM25 baseline for informational intent and its subcategory factual intent, which means that those models still benefit from the semantics of the queries despite the lack of context.

Although our query intent predictor is applicable to all the collections present in TIRA, it might not be very useful for some of them. For the Touche collections that contain non-factoid queries, such as argumentative questions (e.g. "should nuclear weapons be abolished?") it will not provide a classification that will be fine-grained enough. A separate intent classification that labels non-factoid queries such as the one suggested by [14] should be used for those collections. A further analysis to determine the test collections that would benefit the most from ORCAS-I query intent predictor would show how to use it in the best way in TIRA pipeline.

Acknowledgments

This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval – DoSSIER (H2020-EU.1.3.1., ID: 860721) and by the OpenWebSearch.eu project (GA 101070014).

References

[1] A. J. Ratner, C. M. De Sa, S. Wu, D. Selsam, C. Ré, Data programming: Creating large training sets, quickly, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/6709e8d64a5f47269ed5cea9f625f7ab-Paper.pdf.

[2] N. Craswell, D. Campos, B. Mitra, E. Yilmaz, B. Billerbeck, Orcas: 18 million clicked query-document pairs for analyzing search, arXiv preprint arXiv:2006.05324 (2020).

[3] D. Alexander, W. Kusa, A. P. de Vries, ORCAS-I: Queries Annotated with Intent using Weak Supervision, in: SIGIR '22: Proceedings of the 45rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022. URL: https://arxiv.org/abs/2205.00926.

[4] S. Farzana, M. Fröbe, M. Granitzer, G. Hendriksen, D. Hiemstra, M. Potthast, S. Zerhoudi, 1st International Workshop on Open Web Search (WOWS), in: Advances in Information Retrieval. 46th European Conference on IR Research (ECIR 2024), Lecture Notes in Computer Science, Springer, 2024.

[5] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

[6] M. Fröbe, J. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, The Information Retrieval Experiment Platform, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023), ACM, 2023, pp. 2826–2836. URL: https://dl.acm.org/doi/10.1145/3539618.3591888. doi:10.1145/3539618.3591888.

[7] A. Broder, A taxonomy of web search, SIGIR Forum 36 (2002).

[8] B. J. Jansen, D. L. Booth, A. Spink, Determining the informational, navigational, and transactional intent of web queries, Information Processing & Management 44 (2008).

[9] M. Kellar, C. Watters, M. Author, A field study characterizing web-based information seeking tasks, JASIST 58 (2007) 999–1018. doi:10.1002/asi.20590.

[10] J. Kim, Task as a predictable indicator of information seeking behavior on the Web, Ph.D. thesis, Rutgers University, 2006.

[11] W. Kusa, D. Alexander, A. P. de Vries, Orcas-i, 2022. doi:10.48436/pp7xz-n9a06.

[12] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, 2020, pp. 39–48.

[13] R. Nogueira, Z. Jiang, R. Pradeep, J. Lin, Document ranking with a pretrained sequence-to-sequence model, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 708–718. URL: https://aclanthology.org/2020.findings-emnlp.63. doi:10.18653/v1/2020.findings-emnlp.63.

[14] V. Bolotova, V. Blinov, F. Scholer, W. Croft, M. Sanderson, A non-factoid question-answering taxonomy, 2022, pp. 1196–1207. doi:10.1145/3477495.3531926.