Dataset of Natural Language Queries for E-Commerce

Andrea Papenmeier, Dagmar Kern, Daniel Hienert
firstname.lastname@gesis.org
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany

Alfred Sliwa, Ahmet Aker, Norbert Fuhr
firstname.lastname@uni-due.de
University of Duisburg-Essen
Duisburg, Germany

ABSTRACT

Shopping online is more and more frequent in our everyday life. For e-commerce search systems, understanding natural language coming through voice assistants, chatbots or from conversational search is an essential ability to understand what the user really wants. However, evaluation datasets with natural and detailed information needs of product-seekers which could be used for research do not exist. Due to privacy issues and competitive consequences, only few datasets with real user search queries from logs are openly available. In this paper, we present a dataset of 3,540 natural language queries in two domains that describe what users want when searching for a laptop or a jacket of their choice. The dataset contains annotations of vague terms and key facts of 1,754 laptop queries. This dataset opens up a range of research opportunities in the fields of natural language processing and (interactive) information retrieval for product search.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI)*; **Natural language interfaces**; • **Information systems** → **Query intent**; • **Computing methodologies** → *Discourse, dialogue and pragmatics.*

KEYWORDS

Dataset; Natural Language Query; User Intent; E-Commerce.

1 INTRODUCTION

To search for products online is an everyday activity of millions of users, with the market share of e-commerce continually increasing [13]. Understanding natural language in deep is the upcoming key technology for searching in e-commerce and the general Web: Voice assistants rely on processing spoken natural language, chatbots need to extract the user's information need from written natural language, and the research field of conversational search explores a dialogue-driven approach to support finding the right information. Research on query formulation with children shows that natural language is the intuitive way of interacting with search engines [9]. Keyword search, however, is an artefact of the search engine's inability to process open vocabularies and extract the essential key facts from a natural language text. Research on query and information need formulation has mainly built on log data [19, 20, 32] or proprietary data [8]. Log data is not suitable to investigate the natural information need, as the system influences the user, i.e. if users believe the system can only handle keywords, they formulate their query accordingly [10]. Some smaller datasets of natural language information needs exist, e.g. as collected by Kato et al. [10]. In the book domain, The CLEF Social Book Search dataset [11] provides 120 natural language information needs. The data

originates from an online discussion forum, which represents a non-controlled data collection setting. In previous research on a small dataset of 132 natural language queries of laptops, we have shown that natural language queries have the potential to reveal more information about the user's target product than queries issued on current search engines [14]. However, available datasets are not big enough to train automated systems to process natural language information needs.

In this work, we collected and curated a large dataset containing 3,540 natural language queries for two product domains (laptops and jackets). Unlike existing datasets, our product queries were collected in a controlled experiment from participants with a broad range of domain knowledge. For the laptop domain and a subset of the jacket domain, we offer manual annotations of the key product facts mentioned in the descriptions, and vague words contained in the texts. With this dataset, we contribute a valuable resource for the field of natural language processing and interactive information retrieval in the context of product search.

2 RELATED WORK

2.1 Data Collections of Information Needs

One of the most chosen instruments for understanding user needs is the extraction from *query logs*. For example, a rough distinction is made by Broder [4] who defined the user intent of web search queries to be either informational, navigational or transactional. Several research works cluster either search queries from logs [19, 20, 32] or transcribed voice logs [8] into groups of user intent. There is certainly a big difference between the user's natural information need and the short keyword-like queries which can be found in query logs. A more detailed description of information needs is given in *TREC* [29] topics and the works based upon (e.g. [2, 16]). Here, the situation and context of the search intent are described in more detail and formed into background stories or search tasks. The CLEF Social Book Search dataset [11] contains a collection of 120 natural language information needs of books extracted from an online discussion forum. Bogers et al. [3] likewise focus on forums to extract natural language information needs. They annotated 1041 information needs (503 in the book domain, 538 in the movies domain) with respect to "relevance aspects", i.e., requirements for the search target.

Another direction is *conversational search*, where the conversation approach fetches more details on the actual information need of the user. Evaluation datasets base on already existing question-answering interactions as available in forums or dialogue support systems. For example, MS Dialog [17] consists of 2,400 annotated dialogues about Microsoft products, and Penha et al. [15] provide a dataset of 80k conversations across 14 domains which they extracted from Stack Exchange.

Human-human conversations in real as well as in Wizard-of-Oz situations in which humans simulate the system are another source of naturally formulated information needs. For example, the CCPE-M dataset [18] focuses on movie preferences, while the Frames dataset [5] focus on the travelling domain with dialogues gathered in a human-human conversation using SLACK. An additional approach in conversational search is to ask *clarifying questions*. The Qulac dataset [1] collected 10k clarifying questions with answers for 198 TREC topics in a crowdsourcing experiment.

Datasets on *spoken information-seeking conversations* between humans provide audio, transcriptions and additional annotations. Spoken conversations show differences to written conversations and can be used for the evaluation of software agents such as Siri, Google Now, or Cortana. The MISC dataset [25] contains audio, video, transcripts, affectual and physiological signals, computer use and survey data for five different search tasks based on topics from previous research. The SCSdata [26] contains 101 transcribed conversations with annotations and video to solve information needs based on nine search tasks and background stories.

2.2 Product Search in E-commerce

Product search and *e-commerce* is a rather new field in academic research and has specialised challenges for information retrieval: documents, queries, relevance, ranking, recommenders, and user interactions are different from well-known research areas such as Web search [28]. On the level of *user intents*, Su et al. [24] distinguish between three different user goals: target finding, decision making, and exploration, which all have different behaviour patterns of query formulation, browsing, and clicking. Sondi et al. [22] report on another taxonomy of queries generated by clustering queries from a log: shallow exploration, major-item shopping, targeted purchase, minor-item shopping, and hard-choice shopping. *Conversational e-commerce* is a new area where the user conducts a dialogue with the conversational system by voice or chat to find the right product or to get help. The dialogue needs to be natural so that the customer feels engaged. Therefore, understanding the overall intent of the user's request is essential [28].

A current paper on research in e-commerce from the SIGIR Forum [28] lists 28 datasets for e-commerce search and recommendations. However, most of the datasets focus on product catalogues and taxonomies as well as on reviews and recommendations. Only two datasets contain search logs with user queries. Most research works in the area of e-commerce and product search of global players use their own query logs to improve their own search systems (e.g. eBay [27], Amazon [23] or Alibaba [31]), but keep the logs confidential. However, even if publicly available, these datasets contain only keyword queries and not the natural information needs the user is thinking of before entering it into a search bar.

Although the use case of product search is essential in e-commerce, there is little data about the genuine information needs of product buyers. To fill the gap of an openly available and controlled collected research dataset, we present in this work a collection of 3,540 natural language queries in product search.

	Laptop	Jacket
Participants (total)	1818	1818
Valid queries (total)	1754	1786
Age (mean, std)	36 (13)	36 (13)
Gender distribution (m, f, d)	700, 1040, 12	718, 1054, 12
Domain knowledge (mean, std)	4.5 (1.5)	4.6 (1.3)

Table 1: Participants statistics for the laptop and jacket queries dataset.

3 DATASET GENERATION

3.1 Data Collection

We recruited 1,818 participants on the crowdsourcing platform Prolific[1] to participate in the experiment. To avoid the effects of the individual language level on the formulations, participants had to be native English speakers. Furthermore, participants were not allowed to use a mobile phone to complete the survey in order to avoid effects from small screen and keyboard sizes.

After giving informed consent, participants were either asked to describe a laptop (imagining their current laptop broke down recently) or a jacket (imagining they lost their jacket). We define those product descriptions as *natural language queries*. All participants completed both tasks, but in randomised order. The task description and the full questionnaire are available online[2]. Finally, the participants answered questions about their demographic background: (1) their age, (2) their gender, and (3) their self-assessed domain knowledge for both domains (on a scale of 1 = "no knowledge" to 7 = "high/expert knowledge"). Table 1 presents the demographic characteristics of the participants for both domains.

All descriptions were manually filtered to eliminate queries that were invalid due to their form (e.g. empty strings) or their content (e.g. the text described a different product, the text did not describe any product, the text was a meta-comment of the participant about the task). From the 1,818 participants, we curated a dataset containing 1754 valid laptop queries and 1786 valid jacket queries.

3.2 Data Annotation

After collecting the natural language queries, we recruited 20 annotators who were taking part in a seminar at our institution. Each laptop query was annotated by three annotators concerning key facts and vague words. Key facts are words or phrases describing requirements of the product, while vague words are words (within key facts) which are ambiguous and depending on interpretation. From the jacket corpus, so far, a subset of 363 queries was annotated concerning the key facts for cross-domain evaluation.

Before starting the annotation task, annotators received the definition of key facts and vague words, examples, and guidance on how to handle negations and borderline cases for vagueness[3]. Annotators also discussed the guidelines in a plenary session. The

---

[1]https://www.prolific.co

[2]https://git.gesis.org/papenmaa/chiir21_naturallanguagequeries/tree/master/Questionnaire

[3]The annotation material is available at https://git.gesis.org/papenmaa/chiir21_naturallanguagequeries/blob/master/Annotationguideline.pdf

	Laptop
Queries (total)	1754
Words per query (mean, std)	35 (20)
Key facts	
Annotated queries (total)	1752
Annotated words per query (mean, std)	10.6 (6.5)
Inter-annotator agreement	.697
Vague words	
Annotated queries (total)	1686
Annotated words per query (mean, std)	3.2 (2.5)
Inter-annotator agreement	.653

Table 2: Dataset statistics for the laptop corpus.

annotation process was conducted on Doccano[4] as a sequence labelling task. Annotators labelled key facts (consisting of one or more words), e.g. (shown here in bold):

> I would buy a **basic** laptop of **any brand**, one with **good reviews**.

Furthermore, annotators labelled vague words, e.g.(shown here in bold):

> I would buy a **basic** laptop of any brand, one with **good** reviews.

The inter-annotator agreement (Krippendorff's alpha, on word-level) on the laptop queries is .697 for the annotation of key facts and .653 for the annotation of vague words. The annotations of jacket key facts have a mean inter-annotator agreement of .697. We provide the implementation of the agreement measure together with the dataset[5]. Table 2 shows detailed annotation statistics for the laptop domain.

Listing 1 presents a single data point from the laptop domain of the final dataset, containing a unique ID, the domain, the original text (unprocessed), information about the user who wrote the text, and the annotations for both key facts and vague words. Each data point contains the individual annotations as well as a combined annotation showing words that were labelled by at least two annotators. The annotated words are identified by the word itself and its position (character-level offset) in the text. The annotated words of the key facts are also available as text segments, where the individual labels have been taken into account.

3.3 Dataset Availability

The dataset is publicly available[6] under a CC BY-NC-SA 3.0 licence. The repository is hosted by GESIS – Leibniz Institute for the Social Sciences, a well-known data provider for social science data. We provide the dataset in JSONL and CSV format, along with a description of variables and the annotation guidelines. Additionally, we provide a Jupyter Notebook with code to import the dataset into Python, perform basic statistical analyses, calculate the inter-annotator agreement, and access single data points.

---

[4]https://github.com/doccano/doccano
[5]Jupyter Notebook available at https://git.gesis.org/papenmaa/chiir21_naturallanguagequeries/blob/master/IAA.ipynb
[6]https://git.gesis.org/papenmaa/chiir21_naturallanguagequeries

Listing 1: Structure of a single data point in the data set.

```
1  { "ID" : 1887 ,
2    "domain" : 'laptop'
3    "text" : 'I want a laptop primarily for internet
          use, it needs to be light with a long
          battery life.',
4    "user" : {
5      "age" : 47 ,
6      "domain knowledge" : 3 ,
7      "gender" : 'male'
8    },
9    "vague words" : {
10     "annotator1" : [['light',59], ['long',72]],
11     "annotator2" : [['light',59], ['long',72]],
12     "annotator3" : [['light',59], ['long',72]],
13     "annotation_by_2" : [['light',59], ['long',72]
          ],
14     "IAA" : 1.0
15   },
16   "key facts" : {
17     "annotator1" : [['light',59], ['long',72], ['
          battery',77], ['life',85]],
18     "annotator2" : [['internet',30], ['use',39], [
          'light',59], ['long',72], ['battery',77],
          ['life',85]],
19     "annotator3" : [['light',59], ['long',72], ['
          battery',77], ['life',85]],
20     "annotation_by_2" : [['light',59], ['long',72]
          , ['battery',77], ['life',85]],
21     "segments" : ['long battery life', 'light'],
22     "IAA" : 0.8107
23   }
24  }
```

4 USE CASES

The proposed natural language queries dataset can be leveraged for multiple tasks in the field of Natural Language Processing (NLP) and Information Retrieval (IR). The former could use this dataset to understand the semantic intents in natural language queries, while the latter could profit from building up (domain-specific) retrieval models for vague conditions. We discuss potential use cases in the following section.

4.1 Natural Language Processing

4.1.1 Spelling correction. The presented natural language queries have been written by users without any formal restrictions and thus contain many typographical and grammatical errors. The following example from our dataset demonstrates this issue: *"i would buy a lenovo as u can also use rhem as tablets which isvery handy"*. This example contains misspelled terms like *"rhem"* and fusion error terms such as *"isvery"*. Furthermore, entity-specific errors like misspelt brand names, e.g. *"Lesovo"*, and domain-specific slang expressions like *"has at least 8 gigs or ram"* are recurring phenomenons in this dataset. Containing different spelling error types and colloquial language expressions, this dataset calls for correction models in order to proceed with tasks like named entity recognition or information retrieval in product search (cf. [30]). The development of

preprocessing techniques regarding raw natural language queries can be researched by investigating this dataset.

4.1.2 Vagueness. One common problem in information retrieval is the vocabulary mismatch between the user's query language and the system's indexing language [6]. This is due to the vague information needs on the user-side where one is not able to sharpen the borders of different concepts, e.g.:

> *I would like one with good battery and high RAM that boots relatively quickly*

The vagueness problem can increase with a higher lack of domain knowledge [14]. Hence, developing automatised models that are capable of recognising vague phrases in product search are needed.

This dataset is annotated on word-level regarding vague expressions. Classifying such vague words is helpful to distinguish between specific and vague conditions. In the case of specific conditions, a retrieval model can try to match such conditions with retailer-generated product information to filter the results. However, in case of vague query conditions, it is not straightforward to apply such filtering. Therefore, product retrieval systems could use user-generated content of products, e.g. user reviews, to filter for user requirements that are not entailed in the retailer-generated fields, e.g. product quality [12]. In previous work, we demonstrated that user reviews are highly correlating with natural language queries according to lexical matching measurements [14].

4.2 Attribute Mapping

Natural language queries differ to keyword queries according to length, i.e. number of query terms, and the desired amount of conditions a certain object needs to meet. Named Entity Recognition (NER) is one task that aims to identify the different categories in natural language texts and can be applied to search queries [7]. As this dataset has been annotated on key fact-level, i.e. requirements that a product needs to satisfy, it can be used to research automatic matching of unstructured to structured information in product search (cf. [28]). Identifying the attribute domains of these key facts is useful for product retrieval systems as matching procedures can be conducted with the technical fields of products.

As natural language queries are characterised by a more complex structure opposed to keyword queries, another interesting task is to parse their semantic structure. Understanding and representing the meaning is more beneficial than using lexical matching methods like BM25.

4.3 Product Query Classification

This dataset has been annotated for the domain of laptops as well as for the domain of jackets. Product retrieval systems require category identification of a search query before applying the matching models. Misunderstanding the query's domain will lead to dissatisfying results. Product query classification has already been researched in the case of keyword search query data [21, 33]. However, solving product query classification on natural language queries could initiate the investigation of more sophisticated algorithms.

5 CONCLUSION AND FUTURE WORK

Natural language is an interesting challenge in product search. Currently, only few research work has focussed on collecting the unbiased natural information need of search engine users. We provide a dataset of 3,540 natural language queries of laptops and jackets. We annotated 1,754 laptop queries concerning the contained key facts and vague words, and 363 jacket queries.

As this dataset is part of an ongoing research project, we plan to enrich the dataset further. First, we aim to annotate the remainder of the jacket queries with vague words and key facts to have a fully comparable dataset in a second product domain. Secondly, we plan to add clean versions of the queries, corrected for spelling mistakes and punctuation. Thirdly, to enable work on interactive information retrieval and user experience design, the key facts should be matched to structured product attributes. In [14], we have made a first investigation of a smaller dataset (N = 132), where we mapped key facts to facets of existing product search engines and clustered unmatched key facts to determine new facets. To train classifiers on matching key facts to the correct facets, a ground truth is needed which we would like to add to the dataset in the future. Finally, for a subset of the product queries, we aim to add relevant products from a product pool to facilitate retrieval experiments.

For deep learning methods, however, the proposed dataset has a rather small sample size and could be further enlarged. Similarly to previous datasets, the proposed dataset is restricted to two product domains. To facilitate insights into the generalisability of models based on this dataset, more product domains should be added.

REFERENCES

[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 475–484. https://doi.org/10.1145/3331184.3331265

[2] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User Variability and IR System Evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 625–634. https://doi.org/10.1145/2766462.2767728

[3] Toine Bogers, Maria Gäde, Marijn Koolen, Vivien Petras, and Mette Skov. 2018. "What was this Movie About this Chick?". In *Transforming Digital Worlds*, Gobinda Chowdhury, Julie McLeod, Val Gillet, and Peter Willett (Eds.). Springer International Publishing, Cham, 323–334.

[4] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (Sept. 2002), 3–10. https://doi.org/10.1145/792550.792552

[5] Layla El Asri, Hannes Schulz, Shikhar Kr Sarma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 207–219.

[6] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. 1987. The vocabulary problem in human-system communication. *Commun. ACM* 30, 11 (1987), 964–971.

[7] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named Entity Recognition in Query. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 267–274. https://doi.org/10.1145/1571941.1571989

[8] Ido Guy. 2016. Searching by Talking: Analysis of Voice Queries on Mobile Web Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association

for Computing Machinery, New York, NY, USA, 35–44. https://doi.org/10.1145/2911451.2911525

[9] Yvonne Kammerer and Maja Bohnacker. 2012. Children's Web Search with Google: The Effectiveness of Natural Language Queries. In *Proceedings of the 11th International Conference on Interaction Design and Children* (Bremen, Germany) *(IDC '12)*. Association for Computing Machinery, New York, NY, USA, 184–187. https://doi.org/10.1145/2307096.2307121

[10] Makoto P. Kato, Takehiro Yamamoto, Hiroaki Ohshima, and Katsumi Tanaka. 2014. Cognitive Search Intents Hidden behind Queries: A User Study on Query Formulations. In *Proceedings of the 23rd International Conference on World Wide Web* (Seoul, Korea) *(WWW '14 Companion)*. Association for Computing Machinery, New York, NY, USA, 313–314. https://doi.org/10.1145/2567948.2577279

[11] Marijn Koolen, Toine Bogers, Maria Gäde, Mark Hall, Iris Hendrickx, Hugo Huurdeman, Jaap Kamps, Mette Skov, Suzan Verberne, and David Walsh. 2016. Overview of the CLEF 2016 social book search lab. In *International conference of the cross-language evaluation forum for European languages*. Springer, 351–370.

[12] Felipe Moraes, Jie Yang, Rongting Zhang, and Vanessa Murdock. 2020. The Role of Attributes in Product Quality Comparisons. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (Vancouver BC, Canada) *(CHIIR '20)*. Association for Computing Machinery, New York, NY, USA, 253–262. https://doi.org/10.1145/3343413.3377956

[13] U.S. Department of Commerce. 2020. *Quarterly Retail E-Commerce Sales*. Technical Report. U.S. Department of Commerce. https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf Accessed on 22.10.2020.

[14] Andrea Papenmeier, Alfred Sliwa, Dagmar Kern, Daniel Hienert, Ahmet Aker, and Norbert Fuhr. 2020. 'A Modern Up-To-Date Laptop'-Vagueness in Natural Language Queries for Product Search. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 2077–2089.

[15] Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing MANtIS: a novel multi-domain information seeking dialogues dataset. *arXiv preprint arXiv:1912.04639* (2019).

[16] Martin Potthast, Matthias Hagen, Michael Völske, and Benno Stein. [n.d.]. Exploratory Search Missions for TREC Topics. ([n. d.]).

[17] Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and Characterizing User Intent in Information-Seeking Conversations *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 989–992. https://doi.org/10.1145/3209978.3210124

[18] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*.

[19] Eldar Sadikov, Jayant Madhavan, Lu Wang, and Alon Halevy. 2010. Clustering Query Refinements by User Intent. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, North Carolina, USA) *(WWW '10)*. Association for Computing Machinery, New York, NY, USA, 841–850. https://doi.org/10.1145/1772690.1772776

[20] Yelong Shen, Jun Yan, Shuicheng Yan, Lei Ji, Ning Liu, and Zheng Chen. 2011. Sparse Hidden-Dynamics Conditional Random Fields for User Intent Understanding. In *Proceedings of the 20th International Conference on World Wide Web* (Hyderabad, India) *(WWW '11)*. Association for Computing Machinery, New York, NY, USA, 7–16. https://doi.org/10.1145/1963405.1963411

[21] Michael Skinner and Surya Kallumadi. 2019. E-commerce Query Classification Using Product Taxonomy Mapping: A Transfer Learning Approach.. In *eCOM@SIGIR*.

[22] Parikshit Sondhi, Mohit Sharma, Pranam Kolari, and ChengXiang Zhai. 2018. A Taxonomy of Queries for E-Commerce Search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 1245–1248. https://doi.org/10.1145/3209978.3210152

[23] Daria Sorokina and Erick Cantu-Paz. 2016. Amazon search: The joy of ranking products. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 459–460.

[24] Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. 2018. User Intent, Behaviour, and Perceived Satisfaction in Product Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 547–555. https://doi.org/10.1145/3159652.3159714

[25] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*, Vol. 5.

[26] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. How Do People Interact in Conversational Speech-Only Search Tasks: A Preliminary Analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. ACM, New York, NY, USA, 325–328. https://doi.org/10.1145/3020165.3022144

[27] Andrew Trotman, Jon Degenhardt, and Surya Kallumadi. 2017. The Architecture of eBay Search. In *Proceedings of the SIGIR 2017 Workshop On eCommerce co-located with the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, eCOM@SIGIR 2017, Tokyo, Japan, August 11, 2017 (CEUR Workshop Proceedings, Vol. 2311)*, Jon Degenhardt, Surya Kallumadi, Maarten de Rijke, Luo Si, Andrew Trotman, and Yinghui Xu (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-2311/paper_14.pdf

[28] Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. 2020. Challenges and Research Opportunities in eCommerce Search and Recommendations. In *SIGIR Forum*, Vol. 54.

[29] Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press.

[30] Chao Wang and Rongkai Zhao. 2019. Multi-Candidate Ranking Algorithm Based Spell Correction.. In *Proceedings of the SIGIR 2019 Workshop On eCommerce co-located with the 42st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019) Paris, France, July 25, 2019. (CEUR Workshop Proceedings)*, Jon Degenhardt, Surya Kallumadi, Utkarsh Porwal, and Andrew Trotman (Eds.).

[31] Jizhe Wang, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, and Dik Lun Lee. 2018. Billion-scale commodity embedding for e-commerce recommendation in alibaba. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 839–848.

[32] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building Task-Oriented Dialogue Systems for Online Shopping. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, California, USA) *(AAAI'17)*. AAAI Press, 4618–4625.

[33] Hang Yu and Lester Litchfield. 2020. Query Classification with Multi-objective Backoff Optimization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1925–1928.