Human-Based Query Difficulty Prediction

Adrian-Gabriel Chifu[1], Sébastien Déjean[2], Stefano Mizarro[3], and Josiane Mothe[4]

[1] LSIS - UMR 7296 CNRS; Aix-Marseille Université, France
adrian.chifu@lsis.org
[2] Institut de Mathématique, Université Paul Sabatier de Toulouse, France
sebastien.dejean@math.univ-toulouse.fr
[3] University of Udine, Italy
mizzaro@uniud.it
[4] Univ. de Toulouse, ESPE, IRIT UMR5505 CNRS, France
mothe@irit.fr

Abstract. The purpose of an automatic query difficulty predictor is to decide whether an information retrieval system is able to provide the most appropriate answer for a current query. Researchers have investigated many types of automatic query difficulty predictors. These are mostly related to how search engines process queries and documents: they are based on the inner workings of searching/ranking system functions, and therefore they do not provide any really insightful explanation as to the reasons for the difficulty, and they neglect user-oriented aspects. In this paper we study if humans can provide useful explanations, or reasons, of why they think a query will be easy or difficult for a search engine. We run two experiments with variations in the TREC reference collection, the amount of information available about the query, and the method of annotation generation. We examine the correlation between the human prediction, the reasons they provide, the automatic prediction, and the actual system effectiveness. The main findings of this study are twofold. First, we confirm the result of previous studies stating that human predictions correlate only weakly with system effectiveness. Second, and probably more important, after analyzing the reasons given by the annotators we find that: (i) overall, the reasons seem coherent, sensible, and informative; (ii) humans have an accurate picture of some query or term characteristics; and (iii) yet, they cannot reliably predict system/query difficulty.

1 Predicting Query Difficulty

The purpose of a query difficulty predictor is to decide whether an Information Retrieval (IR) system is able to properly answer a current query, that is to say, if it is capable of retrieving only the relevant documents that meet a user's information need as expressed through his or her query. Predicting query difficulty is a hot topic: if a search engine could predict its own chances of failure when processing a given query, it could adapt its processing strategies to increase the overall effectiveness, perhaps even by requesting more information directly from the user in order to better meet his or her needs. The example of an ambiguous term is a textbook case; given the query "Orange", the system can predict that the query may be difficult because the term has various meanings attached to it. The system may decide to diversify its answers to encapsulate the various meanings of the word; or it may ask the user if he or she is interested by the telecom

company, the color, the fruit, or by something else; it may also derive the meaning from the user's past queries, if available. However, ambiguity is not the only reason for a query being difficult (the number of senses of query terms correlates only weakly with system effectiveness [12]).

Predicting query difficulty is challenging. Current automatic predictors are either computed before a search is carried out (pre-retrieval predictors, e.g., the inverse document frequency of the query terms [14]), or computed from a list of retrieved documents (post-retrieval predictors, e.g., the standard deviation between the top-retrieved document scores [13]). The literature reports slightly better correlations with actual system effectiveness when using post-retrieval predictors than pre-retrieval ones, although pre-retrieval predictors are the most interesting for real applications because they are cheaper to calculate. Still, these correlations are weak, even when the various predictors are combined [12,7,6,1]. Moreover, the current automatic predictors are founded on the way search engines process queries and documents, and the way they rank retrieved documents. They do not consider what causes the query to be difficult. Indeed, the features used to calculate automatic predictors are linked to inner functions of search engines, which do not necessarily reflect the human perception of difficulty.

In this paper our intention is to go one step further in query difficulty analysis and understanding, by taking into account the human perspective rather than the system perspective. Instead of considering how IR features could be used to predict difficulty, as it has usually been done so far, we focus on understanding what the *human* perception of query difficulty is and on *why* does a query sound as difficult. To do this, we conduct user studies where we ask human annotators[5] to predict query difficulty and explain the reasons for their prediction. We also aim to understand how different clues on the data and the amount of information provided to the human annotators affect the outcome. To this aim, in some cases the annotators receive only the query as submitted to the system (in our study we consider the *title* field of TREC topics as a query); in other cases the annotators receive, as a surrogate of the user's intent, a longer description of what the user requires (we consider the *descriptive* part of the TREC topic).

Summarize the research questions here according to remarks from reviewers.

The rest of this paper is structured as follows: Section 2 introduces related work; Section 3 discusses the motivations of our approach and frames the research questions that we address. Sections 4, 5, and 6 address each research question. Section 7 mentions future work and concludes the paper.

2 Related Work: Why Queries Are Difficult?

Evaluation in IR has a long history and programs such as TREC have brought many interesting clues on IR processes. One of them is the huge variability in terms of relevance of the retrieved documents, according to both topics and systems.

The Reliable Information Access (RIA) workshop has been the first attempt to try to *understand* in a large scale the IR "black boxes". As stated by Harman and Buckley,

[5] We use the terms "predictors" and "annotators" or "participants" to distinguish between automatic and human prediction, respectively.

"The goal of this workshop was to understand the contributions of both system variability factors and topic variability factors to overall retrieval variability" [4,5]. Harman and Buckley claim that understanding variability in results is difficult because it is due to three types of factors: topic statement, relationship between topics and documents and system features. The RIA workshop focused on the query expansion issue and analyzed both system and topic variability factors on TREC collections. By considering failure analysis, 10 classes of topics were identified manually, but no indications were given on how to automatically assign a topic to a category. One of the main conclusions of the failure analysis was that systems were missing an aspect of the query, generally the same aspect for all the systems. "The other major conclusion was that if a system can realize the problem associated with a given topic, then for well over half the topics studied, current technology should be able to improve results significantly."

Interactive IR studies are somehow related to our work as they involve users and analyze their behavior and difficulties while completing a search task. However, these studies are more oriented on analyzing the users' sessions, their successes and failures.

The most related work is the one from Liu et al. [9,10], that aims at collecting and analyzing why users perceive a given task as difficult. In their study, the users were given complex tasks, such as collecting information to write a new entry in Wikipedia, which is quite different from TREC ad hoc task. Users were asked to provide reasons for pre-task difficulty perception and they mentioned time limitation, complexity and specific requirements. Other aspects were more related to users and interaction which is less related to our work. While Liu et al.'s research focuses on a few search topics and encapsulates the users' knowledge in their schema, we rather consider many search topics and focus on the reason why the system may fail given the query. Thus, we are more in line with RIA workshop, as a system failure analysis project.

Hauff *et al.* [8] analyzed the relationship between user ratings and system predictions using ClueWeb 2009. They study both the topic level and the query level. In the latter, the authors consider various queries for a single topic or information need and measure the users' ability to judge the query suggestion quality. The topic level study is closer to ours: annotators who were provided with the topic title and description were asked to rate the quality of the queries using a five-level scale. The authors found that (i) the inter-annotator agreement is low (Cohen's Kappa between all possible pairs of annotators is between 0.12 and 0.54, (ii) the correlation of individual users and system performance is low (median correlation 0.31 for AP and 0.35 for P@30).

Mizzaro and Mothe [11] carried out a laboratory user study using TREC topics and found that the human prediction only weakly correlates with system performance. That paper also reports some preliminary results on why queries might be perceived as difficult by humans, but the analysis is very limited.

Our study goes a step further: we try to understand the reasons why a query is perceived difficult by analyzing user comments, and we also analyze the relationship between these reasons and human prediction of difficulty as well as with automatic predictors or query features and with actual system performance. We found interesting cues that can be reused in future research either on query difficulty prediction or for improving users' information literacy. Finally, our results suggest that some reasons users can formulate are good predictors of possible system failure.

3 Why Studying Human Query Difficulty Prediction?

In this paper, we go a step further in system failure and query difficulty analysis. Our main goal is to get cues on what users think a difficult query for a system is. These clues may differ from what the system actually finds a difficult query. To this aim, we asked annotators to indicate both their prediction on query difficulty and their explanation for the reason they think the query is going to be easy or difficult for a search engine.

There are several motivations underlying our research and the user study approach that we have chosen. Current understanding of query difficulty and current query difficulty predictors are based on the way queries and documents are processed by the search engine. While we know that tf.idf of query terms has an effect on the system results, we do not know if humans are able to perceive other cues that the systems do not capture, nor if some of the human predictors are correlated to some automatic predictors, giving them more sense to humans. Reversely, it might be that some strategy used by human predictors could be a good automatic predictor, if calculated properly. Also, we do not know yet the theoretical possibilities of automatic query difficulty prediction. By studying how humans predict query difficulty we might be able to understand how difficult the task of automatic prediction is: for example, if predictions based on query terms only are much worse than full information need based predictions (or maybe even impossible at a satisfactory level), then we would have a more precise measure of how difficult (if not impossible) the task of the automatic prediction systems is.

Longer term objectives are: to define pre-retrieval predictors that are based on our findings about human perception and that, hopefully, will be at least as effective as automatic post-retrieval predictors, and better than current pre-retrieval ones; and derive some element for information literacy training. More explicit research questions are:

RQ1 **Difficulty Reasons.** *Why* is a query difficult? Can human annotators identify and express the reasons why a query is difficult? Are these reasons sound? Do these reasons correlate with automatic predictors, and/or with other query features?

RQ2 **Amount of information** Automatic predictors use the query only since they cannot access the user's information need. Do human predictions depend on the amount and kind of information available? Do they evaluate queries in a different way when they know the query only and when they have a more complete description of the user's need?

RQ3 **Links with actual system difficulty.** Are these reasons accurate predictors of perceived or actual query difficulty? Do automatic predictors capture any difficulty reasons called upon by users?

Those research questions (of which RQ1 is probably the most interesting) are addressed by two experiments, named E1 and E2, performed in a laboratory settings and involving users. Overall, the two experiments main features are summarized in Table 1. We varied: the collection used (TREC 6-7-8 and TREC 2014, a.k.a. ClueWeb); the amount of information presented to the user (Query, Q, vs. Query and description, Q+D); and the collected annotations (level of difficulty on a 3 or 5 levels scale; explanation in free text, or explanation through five levels questions/answers). These two slightly different experimental designs allow us to study also two important issues: reasons generation

Table 1. The two experiments. E1 uses TREC ad-hoc collections while E2 use TREC Web track ClueWeb12 collection with TREC 2014 topics. (*) In E1 each participant chose which topics to annotate from the 150 available. (**) Free text annotations were recoded to derive re-coded reasons explaining difficulty, as explained in the text.

	# of Particip.	Scale	Collection	# of topics	Metrics	Amount of info	Explan.	Topics
E1	38 (29 + 9)	3	TREC 6-8	91 (*)	AP	Q, Q+D	Free text (**)	321-350 in TREC 6, 351-381 in TREC 7, 421-450 in TREC 8 (*)
E2	22	5	TREC 2014	25	ERR@20 NDCG@20	Q, Q+D	Categories + Free text (**)	251 255 259 261 267 269 270 273 274 276 277 278 282 284 285 286 287 289 291 292 293 296 297 298 300

(E1) vs. identification (E2); and more longer and complete topic descriptions, but on quite old topics (E1) vs. shorter and less informative topic descriptions, but on more recent topics (E2). More details are presented as needed in the following sections.

4 RQ1: Difficulty Reasons

Our first objective is to know if users can explain why they think a query will be difficult or easy for an engine.

4.1 Finding Reasons: First Experiment (E1)

The first experiment E1 aims to collect free text explanations that participants associate with query ease and difficulty. The participants to E1 were 38 Master's Students (25 1st and 13 2nd year) in library and teaching studies; although they were trained to use specialized search engines, they had just an introduction class on how search systems work. Participants could choose as many topics they wanted to annotate from the set of the 150 topics from TREC 6, 7, and 8 *adhoc* tracks, labeled as 301-450.Each participant was first shown the query only (Q in Table 1 and in the following; it corresponds to the Title part of the TREC topic) and asked to evaluate its difficulty on a three-level scale (Easy / Medium / Difficult), as well as to provide a mandatory explanation in free text. Since the query only might not reflect well the user's intent, the worker was then shown a more complete description of the query (Q+D, i.e., Descriptive and Narrative parts of TREC topic), and the worker again evaluated the query. This two-stages (Q followed by Q+D) prediction was repeated for the queries each participant chose.

Topics were displayed in different order to avoid any bias as the first topics may be treated differently because the task is new for annotators. Moreover, annotators could skip some topics if they wish; this was done to avoid them answering on a topic they did

Table 2. Most frequent: (a) words in free text comments; (b) comments after recoding.

(a)

Easy because		Difficult because	
Precise	113	Missing	64
Clear	48	Broad	62
Many	45	Risk	56
Polysemous	36	Context	34
Usual	16	Polysemous	33
Specialist	15	Vague	26
Simple	11	Many	21

(b)

Easy because		Difficult because	
Precise-Topic	66	Risk-Of-Noise	50
Many-Documents	45	Broad-Topic	43
No-Polysemous-Word	31	Missing-Context	34
Precise-Words	25	Polysemous-Words	22
Clear-Query	19	Several-Aspects	20
Usual-Topic	16	Missing-Where	16

not understand or felt uncomfortable with. The drawback is that the number of annotations varies over topics, and that some topics are not assessed. However, our goal was to collect reasons that humans associate to ease and difficulty, and therefore an association with each topic was not needed. It was instead important to leave the participants free to generate any reason that they might come up with; this is why we used free text. Since the annotation process is difficult, we tried to provide to the students the most favorable conditions. Students could also choose between annotating the query only (and they were not shown the full topic description) or using both the Q part (before) and the full Q+D description (after). Of the 38 students, 29 annotated query difficulty considering Q only, whereas 9 students annotated using both Q and Q+D.

4.2 From Free Text to Re-coded Text

We analyzed difficulty reasons first using free text, then using re-coded free-text.

Manual analysis of free text comments First, we analyzed the free text manually, with the objective of finding if there were some recurrent patterns. When we asked for free text comments we did not provide any comment writing guidance, except from using the keyword "Easy:" or "Difficult:" before any comment. We asked for free text explanations of their query difficulty predictions because we did not want to drive the results. Table 2(a) lists the most frequent words associated with ease and difficulty in the comments. In a few cases, the comments were difficult to understand or analyze because not explicit enough. This was for example the case when annotators wrote *vague* without detailing if it was a query term which they found vague or the topic itself. A typical example is the one of Query 417 from TREC 6-8 (Title:*creativity*) for which the 5 annotators considered the query as difficult using comments such as "too broad, not enough targeted", "far too vague", "far too vague topic", "keyword used very broad, risk of noise", and "a single search term, risk of getting too many results". While some comments are quite explicit, other are difficult to interpret.

Re-coded text Automatic text analysis would have implied to apply advanced natural language processing with no guarantee of success considering, for example, the specificity of the vocabulary, and the lack of data for training. For this reason, we rather

Table 3. Examples of recoding.

Comment	Recoding
A single word in the query	One-Word
The term exploration is polysemous	Polysemous-Word
Far too vague topic	Too-Vague-Topic
Is it in US? Elsewhere?	Missing-Where
Few searches on this topic	Unusual-Topic
Risk of getting too many results	Too-Many-Documents
There are many documents on this	Many-Documents

analyzed manually the free text and re-coded it; which is a common practice in users studies. Table 3 shows some examples of the re-coding we made.

Annotator peculiarities To check the correlation between annotators and the annotations they provide (after re-coding), we used Correspondence Analysis (CA) [2] on the matrix that crosses annotators and re-coded comments (not reported here because of space limits). Compared to more commonly used Principal Component Analysis, CA allows displaying on the same space the variables and observations. We analyzed if some annotators used some specific comments or have different ways of annotating difficulty reasons. We could not find very strong peculiarities among the types of annotations the participants used that would have justified a complementary experiment.

Comments associated to ease and difficulty Table 2(b) displays the most frequent recoded reasons associated to ease (left part) and difficulty (right part). Remember that a given query can be annotated by some comments associated to both. Some phrases are associated both to ease and difficulty of a query (as, e.g., *Many-Documents*). Indeed, users may have in mind recall-oriented tasks and precision-oriented tasks.

While *Precise-Topic* is generally associated to ease (66 times), it is also associated to difficulty in 3 cases. In that cases it is associated to other comments, e.g. *The topic is very precise but it may be too specific*. In the same way, *Many-Documents* is mostly associated to ease and *Too-Many-Documents* to difficulty. When *Many-Documents* is used associated with difficulty, it is generally associated to *Risk-Of-Noise*.

This first analysis helped us in having a better idea on human perception of difficulty. However, E1 was not enough to study real effects because: (i) we had a different number of annotated topics per participants and a different number of annotators per topics; (ii) the free text expression was too hard to analyze; and (iii) the collection was not fully appropriate for humans to annotate query difficulty. We thus designed a second experiment addressing these issues, presented in the next section.

4.3 Reasons as Closed Questions: Second Experiment (E2)

We designed and performed a second experiment E2, with three main differences from E1 (presented in Section 4.1). First, we changed the collection from TREC 6-8 to ClueWeb12. TREC 6-8 collections are widely used and are appropriate for this kind of study, since they feature a large number of topics with a long and detailed descrip-

tion of the needs; these collections are still used for evaluation purposes [15]. But they are old: some participants had difficulties in annotating the queries just because of time reasons (although in the previous setting we made clear in the instructions that the collection contained documents from the 90s). For example in the 90s El Niño was a hot topic in News because it was one of the powerful oscillation events in history, but some of the 2015-16 young students did not hear about the phenomena and event from 1996-97. So in this second experiment we used a newer collection, the ClueWeb12 collection (Category A corpus) used in TREC 2014, which is a large and recent Web snapshot with more recent topics. As a consequence, topics were different too: we selected the 25 topics shown in the third row Table 1, that are the easiest 10, the most difficult 10, and the medium 5 according to the topic difficulty order presented in the TREC track overview paper [3]. One disadvantage of TREC 2014 (that is important to mention because it also justifies the previous experiment) is the rather short query Description.

Second, we switched from three level difficulty to a five-level scale of difficulty ("Very Easy", "Easy", "Average", "Difficult", "Very Difficult") which is more standard.

Third, participants did not express difficulty reasons in free text as in E1, but using closed questions, that we designed on the basis of the free text comments gathered in E1 and of their re-coding. We were able to re-code the comments indicated by the users into reasons phrased as closed questions that could be answered following a scale of values. We used 32 reasons in total (denoted with Ri later on); they are listed in Table 4. We think we cover all the aspects we found in the E1 participants' annotations, i.e., any reason that was expressed in E1 can be expressed also in E2. These reasons were to be answered, both when annotating Q and Q+D, using a five-level scale, ranging from "-2 I strongly disagree" to "$+2$ I strongly agree").

Having the same number of predictions for each query makes the statistical analysis more smooth and sound; we thus consider the same number (8) of annotators for each topic. Participants were 22 volunteers recruited using generic emailing lists mainly from our research institutes, and they got a coupon for participating. Each of them was asked to annotate 10 queries (we took care of using the usual randomized experimental design). Annotators had to annotate the level of difficulty of the query using a five-level scale, but rather than asking to explain the reason of their grading in free text only, we asked them to answer the predefined closed questions. As in E1, Q was presented first, then Q+D. We collected 200 annotations of each type in total, with 8 annotators for each of the 25 topics (we removed annotations when we got more than 8). In the rest of the analysis we average the annotations over the participants for each annotated topic.

4.4 Closed-Reasons Analysis

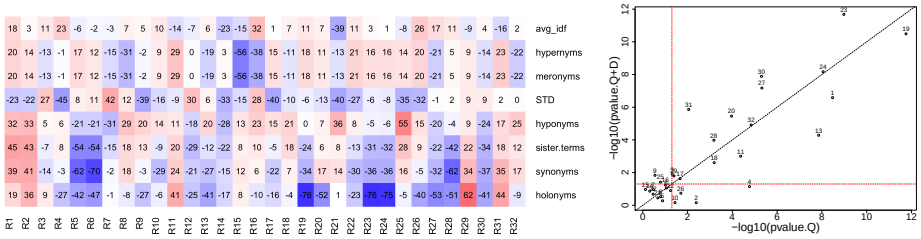
Correlation between human difficulty perception and closed questions/reasons Table 4 shows the correlation between the values humans associate to a reason and the level of difficulty predicted, first when considering Q only, then when considering Q+D. These correlations are obtained after aggregating the results over the 8 annotators and the 25 topics. For example, "R19: None a very few relevant will be retrieved" is strongly correlated to human prediction of query difficulty, as R23 and R24 are, although negatively. Less correlated but still significantly, are R10 (unknown topic), R11 (too broad),

Table 4. Closed reasons resulting from re-coding free text annotations and their correlations with human prediction on Q (column 2) and human prediction on Q+D (column 3).

Reason	**Correlation**	
	Q	Q+D
R1: The query contains vague word(s)	0.523	0.370
R2: The query contains polysemous/ambiguous word(s)	0.342	0.145
R3: The query contains word(s) that is (are) relevant to the topic/query	-0.410	-0.356
R4: The query contains generic word(s)	0.296	0.135
R5: The query contains proper nouns (persons, places, organizations, etc.)	-0.040	0.255
R6: The query contains uncommon word(s)	-0.005	0.024
R7: The query contains specialized word(s)	-0.238	-0.241
R8: The words in the query are inter-related or complementary	-0.028	0.187
R9: The query contains common word(s)	-0.089	0.006
R10: The topic is Unusual/uncommon/unknown	0.526	0.496
R11: The topic is too broad/general/large/vague	0.393	0.502
R12: The topic is specialized	-0.103	-0.136
R13: The topic has several/many aspects	0.614	0.708
R14: The topic is current/hot-topic	-0.118	-0.246
R15: The topic is Non-specialized	-0.017	0.037
R16: The topic is too precise/specific/focused/delimited/clear	-0.149	-0.237
R17: The topic is Usual/common/known	-0.627	-0.512
R18: The number of documents on the topic in the Web/collection is high	-0.693	-0.564
R19: None or very few relevant document will be retrieved	0.880	0.800
R20: Only relevant documents will be retrieved	-0.472	-0.604
R21: There will be different types of relevant documents in the Web/collection	-0.023	0.137
R22: Non-relevant sponsored links/documents will be shown	0.040	0.338
R23: Many of the relevant documents will be retrieved	-0.867	-0.763
R24: Many relevant documents will be retrieved	-0.873	-0.751
R25: Documents with various relevance levels can be retrieved	0.189	0.383
R26: The number of query words is too high	0.624	0.205
R27: The query is concrete/explicit	-0.390	-0.587
R28: The query concerns various aspects	0.458	0.681
R29: The number of query words is too low	0.185	0.353
R30: The query is clear	-0.532	-0.631
R31: The query is missing context	0.273	0.516
R32: The query is broad/vague	0.352	0.615

R13 and R28 (various aspects), R27 (concrete query), R32 (vagueness). All these are interesting reasons that humans relates to difficulty. Other reasons are not correlated with their perception of difficulty such as R5, R6, R9, R12, and R15.

Correlation between closed reasons When we defined the closed questions/reasons from the free texts provided by E1 participants, we tried to avoid to use clearly correlated reasons, but we kept some that were not obviously correlated (e.g."R19: None or very few relevant document will be retrieved", "R23: Many of the relevant documents will be retrieved", and "R24: Many relevant documents will be retrieved"). Nonetheless, it is worth analyzing deeper the correlation between reasons.



(a) (b)

Fig. 1. (a) Correlations (x100) between reasons (X axis) and query features and automatic predictors (Y axis) when using Q. (b) Significance of the correlations between reasons and predicted difficulty (Q on X axis and Q+D on Y Axis).

After having aggregated the data by topic over the annotators, we then calculate the Pearson correlation between reasons. We find that, for example, R1, R2 and R3 strongly correlate, the two first positively while the third negatively. We can also see other groups of correlated reasons: R12 and R15 (a mistake to have kept the two), R10 and R17, R19 and R24. Clearly, not all the reasons are independent, and in future experiments the highly correlated reasons can be removed.

Can human reasons be explained by query features? Some of the closed-reasons are somehow associated with query features used in information retrieval studies. For example, "R2: The query contains polysemous/ambiguous word(s)" can be associated to the number of senses of query terms and thus to the *Synsets* query difficulty predictor, i.e., the number of senses in WordNet (http://wordnet.princeton.edu/) for the query terms proposed by Mothe and Tanguy [12]. "R4: The query contains generic word(s)" can also be associated to linguistic characteristics that could be captured through WordNet. It is thus interesting to check if humans capture these features properly.

We thus analyze the correlation between the reasons and some query features. We consider the Synset linguistic feature which calculates the query terms ambiguity based on WordNet [12]. We add other linguistics features extracted from WordNet. They correspond to the relations that exist between terms in WordNet resource: number of hyperonyms, meronyms, hyponyms, sister terms, synonyms, holonyms. These features were first calculated on each query term, then the median value is kept (we tried min, max, and avg also).We also consider two major statistical features used in the literature as query difficulty predictors: IDF, that measures the fact a term can discriminate relevant from non-relevant documents as a pre-retrieval predictor and STD, the standard deviation between the top-retrieved document scores [13], which is a post-retrieval predictor.

Results are presented in Figure 1(a). The darker the color, the stronger the correlation. First, these results show that the human perception of ambiguity is not very strongly correlated to the ambiguity as capture by WordNet. This point will be worth analyzing deeper in future research that will imply an adhoc user study. On the other hand, the number of query terms WordNet synonyms correlates with R28 (various aspects). That could be a way users express topic ambiguity as well. It also correlates with R6 (specialized word); this make sense since it is likely that specialized words have not

many senses. The number of query terms holonyms (part-of relationships) in WordNet is strongly correlated with human perception of the number of possible relevant documents in the collection. This can be explained by the fact that if a term has a lot of holonyms, it is likely that a lot of documents will exist on the various parts of it.

When analyzing the correlation with IDF and STD automatic query difficulty predictors, we can observe from the figure that there are not very strong correlations with reasons. One of the strongest correlations is between R21 (different types of relevant documents) and IDF in one hand and between STD and R4 (generic words), R7 (specialized words), R17 (Usual/common/known) and STD on the other hand.

5 RQ2: Amount of Information

We now analyze how much the amount of information available to annotators affects both their prediction and annotations. The human prediction on Q and Q+D significantly correlate, although values are never high. The Pearson correlation in E2 for example is of 0.653 with a p-value of $2.2e^{-16}$. They also significantly correlate when using χ^2 considering the annotations as categorical. When moving to TREC 2014, it is always the case that Q is more accurate than Q+D, contrary to our expectations. It seems that the longer description harms, rather than help, in TREC 2014. One possible explanation is that Q+D was much less detailed in TREC 2014 than in TREC 6-8. Also, a psychological effect might have happened: the participants were first shown the short description Q and then, when shown the Q+D, they might have assumed that "something has to be changed", thus worsening their prediction when it was good in first place.

Figure 1(b) reports the statistical significance of the correlation between the closed-reasons and the human prediction of the query difficulty. Each number represents the reason positioned according to its X and Y coordinates. X-axis corresponds to the p-values calculated on annotations collected using Q only while Y-axis corresponds to the p-values calculated when using Q+D, on log scale. The dotted lines (also in red) mark-up the 0.05 significance level. Reasons in the bottom-left rectangle defined by the dotted lines are not significantly correlated with the level of difficulty mentioned by the participants. For example R5 is in that corner; the value a human gives to it does not correlate with his perception of difficulty. On the other hand, reasons in the top-right rectangle are significantly correlated with it, both when considering Q and Q+D annotations (e.g., R19, R23). The value the user gives to the fact that there will be a lot of relevant documents (R23) correlates with his prediction of difficulty. On the bottom-right, the reasons given when considering Q only are significantly correlated with the level of difficulty humanly predicted on Q while the reasons provided on Q+D are not significantly correlated with the difficulty level predicted on Q+D (e.g. R2 and R4). For example "R2: Query contains polysemous words" significantly correlates with the predicted value of difficulty when considering Q but it is no more obvious when considering Q+D. The reverse phenomenon can be observed on the top-left rectangle (e.g., R9). Since we also observed (not reported in detail here due to lack of space) that the values given by the annotators on reasons when considering Q and Q+D highly correlate, what changed here is the perception of difficulty.

6 RQ3: Links with Actual System Difficulty

We also analyze the accuracy of human prediction. For this, we calculate the correlation between human prediction and actual system effectiveness both considering the best system effectiveness for the corresponding TREC track, using the various official measures of the task. For space limits we cannot report detailed results, but all our attempts to detect correlation between human difficulty prediction and system effectiveness have been vain. This result is consistent with the few related work that also focus on this topic [8,11].

7 Conclusion and Future Work

Compared to the RIA workshop [4,5], the annotators for this study are less-specialist in IR. Compared to Liu et al.'s studies [9,10], our study focuses on predicting query difficulty based on the query statement or on the intents of the user, but independently of the user's knowledge on the topic; even though it may have an influence on the annotation they provided. Compared to Hauff et al.'s work [8], we went a step further to understand the users' point of view on query difficulty.

When asking for free text reasons, we found that, overall, the reasons annotators provided seem coherent, sensible, and informative. Moreover, humans have an accurate picture of some query or term characteristics; for example regarding the ambiguity of terms, even if their perception of ambiguity is probably broader than what a linguistic resource can gather. But we also found that humans are bad to predict the difficulty a system will have to answer properly to a query. This result is consistent with the literature. Finally, we found that some reasons they answered through closed-questions are better correlated to actual system effectiveness than automatic predictors from the literature, opening new tracks for research on helping users to formulate their queries.

References

1. S. Bashir. Combining pre-retrieval query quality predictors using genetic programming. *Applied intelligence*, 40(3):525–535, 2014.
2. J.-P. Benzécri et al. *Correspondence analysis handbook*. Marcel Dekker New York, 1992.
3. K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. Voorhees. TREC 2014 Web Track Overview. In *Text REtrieval Conference*. NIST, 2015.
4. D. Harman and C. Buckley. The NRRC reliable information access (RIA) workshop. In *SIGIR*, pages 528–529, 2004.
5. D. Harman and C. Buckley. Overview of the reliable information access workshop. *Information Retrieval*, 12(6):615–641, 2009.
6. C. Hauff. *Predicting the Effectiveness of Queries and Retrieval Systems*. PhD thesis, 2010.
7. C. Hauff, D. Hiemstra, and F. de Jong. A survey of pre-retrieval query performance predictors. In *CIKM*, pages 1419–1420, 2008.
8. C. Hauff, D. Kelly, and L. Azzopardi. A comparison of user and system query performance predictions. In *Conf. on Inf. and knowledge management, CIKM*, pages 979–988, 2010.
9. J. Liu and C. S. Kim. Why do users perceive search tasks as difficult? Exploring difficulty in different task types. In *Symposium on Human-Computer Interaction and Information Retrieval*, 13, pages 5:1–5:10, 2013.

10. J. Liu, C. S. Kim, and C. Creel. Exploring search task difficulty reasons in different task types and user knowledge groups. *Information Processing & Management*, 2014.
11. S. Mizzaro and J. Mothe. Why do you think this query is difficult? a user study on human query prediction. In *SIGIR*, pages 1073–1076. ACM, 2016.
12. J. Mothe and L. Tanguy. Linguistic features to predict query difficulty. In *Predicting query difficulty Wkp, Conf. on Research and Development in IR, SIGIR*, pages 7–10, 2005.
13. A. Shtok, O. Kurland, and D. Carmel. Predicting Query Performance by Query-Drift Estimation. In *ICTIR*, volume 5766 of *LNCS*, pages 305–312, 2009.
14. K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.
15. L. Tan and C. L. Clarke. A family of rank similarity measures based on maximized effectiveness difference. *arXiv preprint arXiv:1408.3587*, 2014.