Comparative Web Search Questions

Alexander Bondarenko[1] Pavel Braslavski[2,3,4] Michael Völske[5] Rami Aly[6]
Maik Fröbe[1] Alexander Panchenko[7] Chris Biemann[8] Benno Stein[5] Matthias Hagen[1]

[1]Martin-Luther-Universität Halle-Wittenberg, [2]Ural Federal University, [3]National Research University Higher School of Economics, [4]JetBrains Research, [5]Bauhaus-Universität Weimar, [6]University of Cambridge, [7]Skolkovo Institute of Science and Technology, [8]Universität Hamburg

{alexander.bondarenko,matthias.hagen}@informatik.uni-halle.de

ABSTRACT

We analyze comparative questions, i.e., questions asking to compare different items, that were submitted to Yandex in 2012. Responses to such questions might be quite different from the simple "ten blue links" and could, for example, aggregate pros and cons of the different options as direct answers. However, changing the result presentation is an intricate decision such that the classification of comparative questions forms a highly precision-oriented task.

From a year-long Yandex log, we annotate a random sample of 50,000 questions; 2.8% of which are comparative. For these annotated questions, we develop a precision-oriented classifier by combining carefully hand-crafted lexico-syntactic rules with feature-based and neural approaches—achieving a recall of 0.6 at a perfect precision of 1.0. After running the classifier on the full year log (on average, there is at least one comparative question per second), we analyze 6,250 comparative questions using more fine-grained subclasses (e.g., should the answer be a "simple" fact or rather a more verbose argument) for which individual classifiers are trained. An important insight is that more than 65% of the comparative questions demand argumentation and opinions, i.e., reliable direct answers to comparative questions require more than the facts from a search engine's knowledge graph.

In addition, we present a qualitative analysis of the underlying comparative information needs (separated into 14 categories like *consumer electronics* or *health*), their seasonal dynamics, and possible answers from community question answering platforms.

KEYWORDS

Question answering; Question classification; Query log analysis

ACM Reference Format:
Alexander Bondarenko, Pavel Braslavski, Michael Völske, Rami Aly, Maik Fröbe, Alexander Panchenko, Chris Biemann, Benno Stein, and Matthias Hagen. 2020. Comparative Web Search Questions. In *The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM '20), February 3–7, 2020, Houston, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3336191.3371848

WSDM '20, February 3–7, 2020, Houston, TX, USA

ACM ISBN 978-1-4503-6822-3/20/02…$15.00
https://doi.org/10.1145/3336191.3371848

1 INTRODUCTION

We permanently face a variety of choices: Where to go for dinner? Which programming language to use? Whether to buy an electric car? In many cases the respective decisions are made by comparing options. A lot of comparison requests can be found on community question answering platforms (CQA) like Yahoo! Answers, Quora, or StackExchange, but also as queries submitted to search engines. More and more of the comparative search engine queries are also formulated as actual natural language questions, a trend that is evoked by the recent advances in speech recognition and the spread of voice interfaces, which encourage users to shift from the telegram-style keyword-based queries to natural language questions [15, 30, 42].

Still, today's web search engines do not treat requests for comparison that are issued as questions any differently to other queries but simply output "ten blue links", regardless of the comparison intent. This misses, for instance, the opportunity to switch the output straight to a direct answer aggregating pros and cons of the different options—similar to the decades-old idea of structured representations in e-commerce [36].

An early proposed solution for comparative information needs was "comparative web search" [39]: to submit each item as a separate keyword query to compare the results. Recently, a slightly more sophisticated search system to tackle comparative information needs was proposed by Schildwächter et al. [33]. However, the system cannot process comparative questions but expects the user to enter the options to be compared along with the comparison aspects in individual fields. An important step towards actually showing a pro/con result presentation for comparative *questions* would be their identification and the study of the underlying information needs. In this paper, we take this step.

Our first contribution is the manual annotation of comparative information needs in a 50,000 question sample from a year-long Yandex log. Four native Russian speaking annotators have labeled the questions as being comparative or not and assigned ten fine-grained subclasses to the comparative ones (e.g., whether a question asks for facts or arguments, whether a superlative is contained, etc.). About 2.8% of the annotated questions are comparative—a non-negligible amount that justifies some deeper inspection.

As our second contribution, we build classifiers for comparative questions and the subclasses using a combination of carefully hand-crafted high-precision rules with feature-based and neural classifiers. The ensemble classifier recalls 60% of the comparative questions with a perfect precision of 1.0. A classifier of that quality is actually applicable in production systems since there are hardly

any false positives to be expected (i.e., almost no wrong switch to a pro/con answer presentation for a question that is not comparative).

Our third contribution is an analysis of the comparative questions detected by the perfect-precision ensemble on the entire 1.5 billion questions from the year-long Yandex log. Many comparative questions fall in the category of consumer electronics (e.g., `Which camera is better, Canon or Nikon?`) followed by cars and transportation (e.g., `Which tires are best for the winter?`). A substantial portion of the frequently asked comparative questions do not specify concrete objects to be compared, and no comparison aspect is provided (e.g., `Which tablet is the best to buy?`). Such queries require more explanatory answers in the form of opinions or pro/con arguments not typically found in a search engine's fact-oriented knowledge graph. We thus also conduct a pilot study to analyze whether answers on similar questions from the Russian question answering platform Otvety can help. For about 50% of the comparative Yandex questions we find a fitting answer on Otvety; in particular, answers for the more frequent comparative Yandex questions are usually "mineable" from the CQA. This potential of mining answers together with our proposed high-precision classification of comparative questions indicates a very promising first step towards handling the result presentation for comparative questions differently than showing "ten blue links" only.

To ensure some level of reproducibility and a potential transfer to English, we release annotations of 15,000 questions from publicly available datasets (Quora [16], MS MARCO [29], and Google Natural Questions [24]) along with our code and pre-trained models.[1] Since we are bound by an agreement, we cannot disclose the Yandex questions themselves.

2 RELATED WORK

Traditionally, comparatives have been considered in linguistic studies [3, 4, 7, 37, 38] as a limited set of lexical structures like comparative adjectives and adverbs or comparative operators (e.g., same–as or different–than). But even though comparative questions are distinguished in a separate class of some question taxonomies [8, 25] the main focus of identifying comparative structures has been on comparative sentences in the field of sentiment analysis. For instance, Jindal and Liu [19] propose a recall-oriented approach of 83 manual rules (keywords and phrases) to identify as many comparative sentences, comparable items, and compared features from reviews as possible (recall of 0.94 at a precision of only 0.34). In an improved variant of their approach, Jindal and Liu [18] train a Naïve Bayes classifier on the manual rules and some learned sequential rules (keywords + POS tags) to achieve a recall of 0.81 at a precision of 0.81. In our approach for comparative question classification, we will also combine rules with other classifiers but with a clear focus on almost perfect precision instead of high recall (we achieve a recall of 0.6 at a precision of about 1.0).

In the field of web search, Jain and Pantel [17] suggested that search engines could better support users with comparative information needs. The proposed approach uses rather simplistic rules like "X vs. Y" to identify comparative keyword queries that are then matched against a look-up table of pairs of comparable items mined from 100 million queries and 500 million web pages. Since comparative questions can use a much richer vocabulary than just simple "X vs. Y"-style patterns, the approach of Jain and Pantel [17] is not directly applicable to our question query scenario but will be adapted using more sophisticated classification steps.

In a study very related to ours, Li et al. [26] proposed a rule-based method for identifying comparative questions (sequential patterns over words, POS tags, placeholders for compared objects, and beginning/end-of-question markers). In an evaluation on 5,200 questions from Yahoo! Answers (about 2.7% comparative), the rules achieved a recall of 0.82 at a precision of 0.83. However, since Li et al. [26] only consider questions that explicitly include two compared items, we cannot use their approach directly (less than 45% of the comparative questions in our sample actually do explicitly contain two items). Instead, we will employ "more general" rules and combine them with other non-rule-based approaches to identify more comparative questions.

3 DATA

To study real-world comparative questions, we mine them from two sources: (1) a year-long log of questions submitted to the Russian search engine Yandex in 2012, and (2) all the questions posted on the Russian question answering platform Otvety in 2012.

From the Yandex query log, we extracted about 2 billion question-style entries that match any of 58 syntactic question indicators (e.g., `how, what, where, should`) similar to the method proposed by Bendersky and Croft [2]—but adapted to Russian. We clean the initial set of about 2 billion entries following the steps of Völske et al. [40]. We remove spam and bot entries by considering a user to be a bot when one of the following conditions holds: (1) more than 2,000 question-style entries over the year, (2) more than 5 question-style entries submitted within any one-minute window, (3) an average length of more than 20 words for the question-style entries, or (4) at least 50 question-style entries in total with the same leading 15 characters in at least 80% of them (e.g., `what is the translation of`). We also remove consecutive duplicate entries from the same user, as well as entries not representing "genuine" user questions (e.g., crossword questions, questions from the TV game show Family Feud, or questions matching Wikipedia titles). These cleansing steps removed about 500 million of the 2 billion question-style entries, resulting in a cleaned set of 1.5 billion entries we consider to be genuine questions (752 million unique questions from 183 million unique user IDs). Interestingly, even though the questions are in Russian, quite many of them contain Latin-spelled tokens (e.g., brands or asking for the correct spelling of some English word).

Following Völske et al. [40] again, we extracted those about 6.6 million questions from the about 11 million questions posted on the Russian community question answering platform Otvety in 2012, for which a best answer was selected and that were asked by the users who posted at least three questions in 2012. Otvety ("answers") is the Russian counterpart of Yahoo! Answers, with similar rules and incentives (points for good answers, etc.). Before posted, each question is manually assigned to one of 28 top-level categories by the asker. In our extraction, we omitted ambiguous categories (humor, miscellaneous, etc.) and merged closely related ones to 14 top-level categories (cf. Table 7).

[1] github.com/webis-de/WSDM-20

Table 1: Absolute and relative frequencies of the comparative question subclasses (percentages for subclasses are relative to the number of comparative questions).

	Yandex		Otvety	
Comparative	**1,405 (3% of all)**		**1,571 (13% of all)**	
Opinion	916	(65%)	1,469	(94%)
Argumentative	676	(48%)	586	(37%)
Reason	83	(6%)	10	(<1%)
Factoid	378	(27%)	101	(6%)
Method	106	(8%)	41	(3%)
Superlative	180	(13%)	287	(18%)
Direct	603	(43%)	893	(57%)
Aspect	302	(22%)	546	(35%)
Context	238	(17%)	405	(26%)
Preference (requested)	985	(70%)	1,281	(82%)
(stated)	18	(1%)	77	(5%)

To ensure a natural distribution of comparative questions, we randomly sampled 50,000 questions that at least three different users submitted from the cleaned Yandex log (these questions are probably less privacy-sensitive), as well as 12,500 questions from the Otvety data. Four native Russian-speaking annotators were told to label as comparative those questions that exhibit an intent of comparing through an examination of (dis-)similarities of two or more items, two or more groups of items, all items inside one group, or a single item against a group of items. The compared items may either be explicitly mentioned (e.g., `Which is better to buy, an iPhone or a Samsung?`) or may be given as a generic "set" (e.g., `Which tablet is best to buy?`). In an initial κ-test on 200 questions, the four annotators reached an inter-annotator agreement of a Fleiss' κ=0.88 (almost perfect agreement) after a round of instructions. Due to the high agreement, the annotators then labeled individual shares of the data independently (i.e., just one vote per question).

Despite extensive automatic pre-filtering to remove non-genuine questions, our annotators still marked about 2,000 of the 50,000 Yandex questions and about 2,500 of the 12,500 Otvety questions as being incomplete, parts of song lyrics our filters missed, or containing profanity. We replaced such questions by additional randomly-sampled questions to maintain the desired totals. Overall, the annotators labeled 1,405 Yandex questions (about 2.8%) and 1,571 Otvety questions (about 12.6%) as comparative.

In a second round of annotations, the annotators then labeled the comparative questions from the first round with the following ten more fine-grained subclasses (annotators achieved a Fleiss' κ of 0.51, a moderate agreement) that are not mutually exclusive (i.e., a question can fall in more than one of the respective classes and the annotators were instructed to select any that applied).

The first five subclasses are general question classes from the literature: *Opinion* questions ask for a personal experience or opinion in the answer without the need of a shared settled knowledge (e.g., `Which to choose for vacation, Goa or UAE?`) [22, 35, 45]. *Argumentative* questions request a solid argumentation in the answer (e.g., `Who will win a presidential election, Trump or Clinton and why?`). *Reason* questions seek an explanation or reasons in the answer that are based on scientific insights (e.g., `What is common between proteins and amino acids?`) [27]. *Factoid* questions can be answered with a simple (often short) fact, where the answer is rather "static" (not changeable) over a sufficient period of time and independent of the answerer's opinion or experience (e.g., `Which contain more vitamin c, kiwis or lemons?`) [1, 28]. *Method* questions request some *how to*-style instruction (e.g., `How to distinguish faux fur from real?`) [27].

In addition to these above five general question classes, we also asked the annotators to assign five further labels focused on syntactic or semantic properties of comparative questions. A comparative question is *superlative* if it asks for the best item in a class (e.g., `Who is the best soccer player?`), rather than explicitly comparing two or more items (e.g., `Who is a better soccer player, Messi or Ronaldo?`). A comparative question is *direct* if it explicitly includes the compared items (e.g., `Which is more reliable, an iPhone or a Samsung?`) instead of implicitly determining a set of the possible items to compare (e.g., `Which mobile phone is it better to buy for 20,000 rubles?`). A comparative question includes an *aspect* when a particular shared property over which the items can be compared or contrasted is mentioned. Such aspects can be stated in ascending or descending direction (e.g., asking whether a product is `more expensive` or `cheaper`), and they can be expressed through a simple comparative adjective or adverb (e.g., `Which is cheaper, an iPhone or a Samsung?`) or through the combination of several lexical units (e.g., `Which is better for web development, PHP or Python?`). Comparative questions may also include additional *context* for the comparison (e.g., target of a 4-year-old in `Which is better to buy for a 4-year-old, a remote control car or a toy transformer?`). Finally, comparative questions may fall in the *preference* class by either *requesting* a preference (e.g., `Which is more reliable, an iPhone or a Samsung?`) or by explicitly *stating* a preference (e.g., `Why is an iPhone better than a Samsung?`).

Table 1 shows the annotation results. Unsurprisingly, users on Otvety (the community question answering platform) ask for relatively way more opinionated comparisons and fewer factoid comparisons than on Yandex. Still, more than 65% of the comparative questions submitted to Yandex also are non-factoid (i.e., directly answering them may be a difficult task).

4 IDENTIFYING COMPARATIVE QUESTIONS

With the scenario of changing a search engine's result presentation for comparative questions in mind, we focus on the precision of classifying comparative questions (about 2.8% of the Yandex questions). We combine three different "techniques" into an ensemble classifier: (1) hand-crafted lexico-syntactic rules, (2) traditional feature-based classifiers, and (3) neural networks. For developing the rules or to train the classifiers, we split the annotated data in training (80%) and test sets (20%).

Rule-based classification. Inspired by previous studies on identifying comparatives [3, 4, 7, 17–19, 26, 37, 38], we use lexical and syntactic rules as a first step of our classifier aiming for perfect precision at a recall as high as possible. We translated promising patterns from the literature to Russian and merged and "tuned" the rules on the training set to not end up with a too large number. Our potential 15 rules below consist of regular expressions over question tokens, comparative (COMP) and superlative (SUPER) grammemes

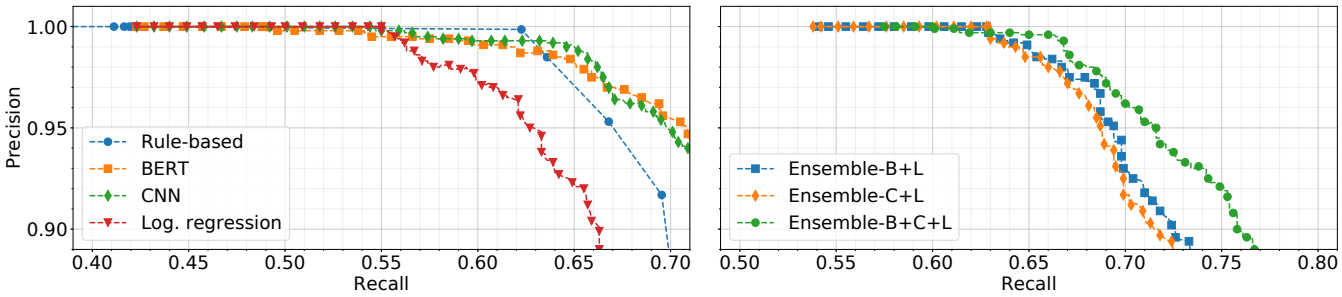

Figure 1: Precision-recall curves for the comparative question class on the Yandex training set (ten-fold cross-validation).

(using the MyStem POS tagger [34]), token positions (posn), and logical operators, and are ordered by descending precision (the ones with equal precision are ordered by descending recall).

(R1) `[better]` ∧ ¬`[how]`[2]
(R2) `COMP` ∧ `[or|vs|versus]`[3] ∧ `posn(COMP) < posn[or|vs|versus]` ∧ ¬`[more or less]`
(R3) `[how correct(ly)? (spell|write)]` ∧ `[or]`
(R4) `[what common|similar]` ∧ `[and|from|or|between|vs|versus]`
(R5) `[choose|buy|take]` ∧ `[or|between|vs|versus]`
(R6) `[in comparison]`
(R7) `[advantage|disadvantage|flaws]` ∧ `[of|over|compared to]`
(R8) `[difference(s)?|differentiate|distinguish]` ∧ `[and|from|or|between|vs|versus]`
(R9) `[better]`
(R10) `COMP` ∧ `[which]` ∧ ¬`[or|vs|versus]` ∧ ¬`[how]`
(R11) `[or]`
(R12) `COMP`
(R13) `COMP` ∧ `[which]` ∧ ¬`[or|vs|versus]`
(R14) `SUPER`
(R15) `[plus(es)?]` ∧ `[minus(es)?]`

Using rules for the classification, a given question will be classified as comparative if any of these rules matches (ignoring punctuation and capitalization). To determine a subset of rules that reach perfect precision, we examine their performance on the training set. The blue line in Figure 1 (left) shows the precision-recall curve resulting from successively adding rules in the descending precision order from above; rules (R1–7) have a perfect precision of 1.0, and together achieve a recall of 0.42. Adding rule (R8) increases recall to 0.62 but slightly reduces precision to 0.9986 (a single misclassified example: `How to teach a dog to distinguish between friends and foes?`). The next rules then provide additional recall but at a much higher cost in precision (e.g., adding the rules (R9–12) increases recall above 0.70 while dropping precision to 0.74). Identifying some more "perfect precision" rules might thus be an interesting direction for future work.

Combination with feature-based and neural classifiers. To supplement the handcrafted rules, we try to add models with manual feature engineering (SVM, logistic regression, Naïve Bayes), as well as neural models (CNN [20], LSTM with recurrent dropout [14], capsule networks with dynamic routing [32] similar to the CapsNet-1 model [43], and BERT with a linear layer as a decoder on top [12]), which have become prevalent for text classification tasks. For all these models, we optimize the parameters in a grid search of commonly used value ranges and evaluate the performance in pilot experiments to identify promising combinations.

Since we consider the classification of comparative questions as a highly precision-oriented task, we aim to further increase the recall of the rule-based approach at the smallest possible cost in precision. We thus train and test the feature-based and neural classifiers only on the "more difficult" questions that are not already identified as being comparative by the perfect-precision rule set (R1–7). From the Yandex questions, this leaves 39,524 questions (650 comparative) as the reduced training and 9,876 questions (159 comparative) as the reduced test set. In a pre-processing, each question is tokenized, lowercased, POS-tagged, and punctuation is removed. For the feature-based classifiers we derive unigram bag-of-words representations (SVM, Naïve Bayes) or uni- to four-gram bag-of-words representations (logistic regression) as these are the best performing setups in our pilot experiments. For CNN, LSTM, and capsule networks, fastText embeddings trained on the Russian Wikipedia are used [6]. For BERT, we fine-tune the pretrained bert-based-multilingual-uncased model with WordPiece embeddings as proposed by Devlin et al. [12].

The BERT, CNN, and logistic regression models vastly outperform the other classifiers in our pilot experiments (higher recall at perfect precision). We thus only consider BERT, CNN, and logistic regression as potential add-ons to the hand-crafted rules. The models' hyperparameters are optimized to achieve the highest precision for the comparative question class using grid search and ten-fold cross-validation on the training set.[4] Our proposed "ensemble" of rules with feature-based and neural classifiers is a four-step decision process (pseudo code in Algorithm 1). Given a question q and a set of classification models C (some subset of BERT, CNN, and logistic regression in our case), the ensemble first applies the "perfect precision" rule set (R1–7). Only if these rules do not classify q as

[2]Expressions in `[]` are in regular expression syntax: so a question matching (R1) must contain the token *better* but not the token *how* (tokens being approximate translations from Russian).

[3]Even though *vs* and *versus* are no Russian comparison words, they occasionally occur as such in queries; still, we do not consider them as standalone comparison indicators since only very few questions contain them (< 0.005%), and since a significant number of *vs*-questions are non-comparative (e.g., `What is vs/versus?`).

[4]BERT and CNN use the Adam optimizer [21] and a minibatch size of 32. BERT fine-tuning: hidden units 768, dropout prob. 0.1, learning rate 0.00002, epochs 3, sequence length 128; CNN: filters 25, windows {3,4,5}, learning rate 0.0005, epochs 3, dropout prob. 0.5, loss function: binary cross-entropy loss, sequence length 15. Logistic regression: penalty='l2', solver='liblinear', C=0.01.

Input: question q, classifiers $C \subseteq$ {CNN, BERT, Logistic}
Output: 1 if q is comparative, 0 otherwise
begin
  // Step 1: ''perfect precision'' rule set
  if ruleDecision((R1–7), q) = 1 **then return** *1*;
  // Step 2: ''perfect precision'' classifiers
  foreach $c \in C$ **do**
    if classifierDecision(c, q, perfectPrecisionThreshold(c)) = 1 **then return** *1*;
  end
  // Step 3: consensus of ''non-perfect'' classifiers
  D_C := [classifierDecision(c, q, decisionThreshold(c)) : $c \in C$]
  if unanimous(D_C) **then return** $D_C[0]$;
  // Step 4: almost ''perfect precision'' rule R8
  return ruleDecision((R8), q)
end

Algorithm 1: Pseudo code of our Ensemble-C classifier.

Table 2: Classification results of a ten-fold cross-validation on the training set aiming for the maximal recall at a precision of 1.0 on the comparative class (decision threshold in brackets). All classifiers achieve at least 0.98 precision and 1.0 recall for the non-comparative class.

Individ. model	Recall	F1	Ensembles	Recall	F1
Logistic (0.418)	0.55	0.71	Ens.-B+L (0.632)	0.63	0.77
CNN (0.99447)	0.55	0.71	Ens.-C+L (0.418)	0.63	0.77
BERT (0.99766)	0.49	0.66	Ens.-B+C+L (0.99447)	0.60	0.75

comparative, the models from C are run to classify q (with a decision threshold optimized for perfect precision on the training set). If none of these models classifies q as comparative, the third step asks whether there is a consensus among the classifiers in C at relaxed decision thresholds (aiming for a combined best possible precision tuned on the training set). If these relaxed-threshold classifiers do not reach an unanimous consensus of q being comparative or not, the fourth step just takes the decision of the high-recall but slightly imperfect-precision rule (R8) (precision of 0.9986).

Varying the decision threshold of each classifier from 0 to 1 in Step 3 of the ensemble approach, three ensemble variants performed particularly well in the pilot experiments: (1) Ensemble-B+L with $C = \{\text{BERT, Logistic}\}$, (2) Ensemble-C+L with $C = \{\text{CNN, Logistic}\}$, and (3) Ensemble-B+C+L with $C = \{\text{BERT, CNN, Logistic}\}$. The precision-recall curves for the complete four-step ensembles on the Yandex training set are shown in Figure 1 (right), the parameter settings and recall values for the individual perfect-precision classifiers and for the complete ensembles are given in Table 2. The Ensemble-B+L and Ensemble-C+L outperformed all other classification models on the training set, achieving a recall of 0.63. Since Ensemble-C+L has a much better run time per to-be-classified question, we will use Ensemble-C+L for the run on the year-long Yandex log in our later experiments (cf. Sections 5 and 6).

5 FINE-GRAINED CLASSIFICATION OF COMPARATIVE QUESTIONS

The ten subclasses of comparative questions that our annotators labeled are meant to differentiate several types of comparative intents (cf. Section 3). These types help to better "understand" comparative intents and to decide what and how an answer should be presented [31]. For instance, answers to factoid and probably also many reason questions (`What is common between proteins and amino acids?`) can possibly be found in knowledge bases and can be presented on a result page as a short direct answer [10]. By contrast, answering opinionated and argumentative questions (`Which one to buy, an iPhone or a Samsung and why?`) may trigger a search for fitting (answered) questions on some question answering platform (CQA) or a search for multiple evidences via multi-hop question answering [9, 11, 13] with summaries stemming from several documents [46]. Finally, an answer to a method question (`How to distinguish faux fur from real?`) might also be found on question answering platforms [41] or in how-to collections and will most likely be presented as step-wise instructions.

Interestingly, not many studies have focused on answering comparative questions up to now. The existing studies [33, 36, 39] deal with queries (not questions) where users explicitly provide two items to be compared. This is similar to questions that we call direct comparisons (`Who is the best soccer player, Messi or Ronaldo?`) but the other subclasses besides direct comparisons are usually "ignored" in the previous studies. Our fine-grained categorization aims to close this gap. For instance, superlative questions (`Who is the best soccer player?`) ask for a search over a group of all possible items (all soccer players) in order to find a single superior one. Sometimes, an aspect for the comparison could be explicitly stated in the question (`Who is the best soccer player when it comes to goals scored?`), or not be mentioned which then requires some "guess work" at search engine side. In addition, context like `for a 4-year-old` as part of a comparative question like `Which is better to buy for a 4-year-old, a remote control car or a soccer?` can also further guide the search for an answer. Finally, a preference in a comparative question (or the absence of a preference) indicates whether the answer should explicitly mention some particular item along with a justification (`Which one to buy, an iPhone or a Samsung and why?`) or whether providing several options along with a comparison of their characteristics is preferred (`What are the main differences between mobile phones?`).

Enlarging the set of comparative questions. The manually labeled 50,000 Yandex questions contain 1,405 comparative questions with some of the subclasses containing only few questions (see Table 1 for the subclass distribution). To have a larger training set for neural approaches to classify the fine-grained subclasses, we thus decided to collect additional comparative questions from the Yandex log. In particular, we choose Ensemble-C+L to classify the whole 1.5 billion questions since it was the fastest and most accurate classifier in our experiments on the test set of the 10,000 labeled questions (approximately a few hours of run time vs. several days when BERT is included to classify the entire Yandex log).

From the questions labeled as comparative by Ensemble-C+L in the year-long Yandex question log, our annotators manually annotated another random 5,000 questions into the ten fine-grained

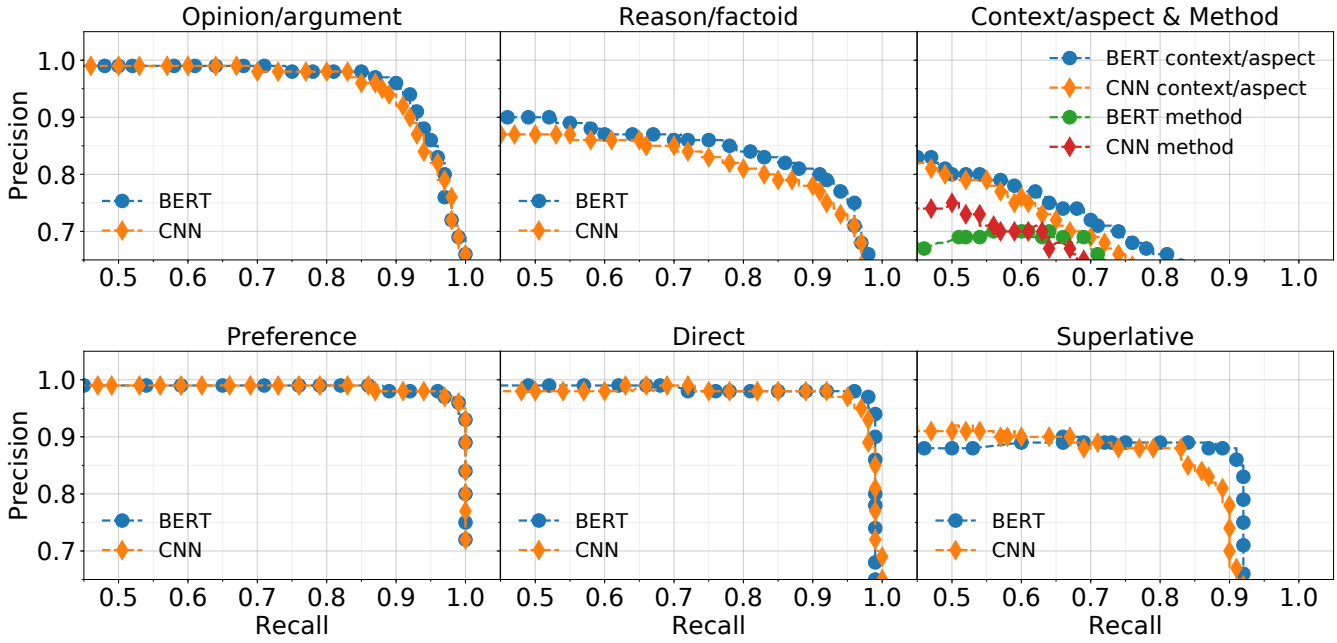

Figure 2: Precision-recall curves for the comparative subclasses on the 5,000 questions training set (ten-fold cross-validation).

subclasses. Since again some questions were labeled as inappropriate by our annotators and since about 1% of the questions actually were not comparative (an expected small number of misclassifications), we ended up with a total of 6,250 comparative questions labeled with fine-grained subclasses (80/20 train/test split). Since some classes were mentioned as closely related by our annotators (in fact, the annotators could assign multiple classes and some classes then overlapped to a large extent), we merge the closely related subclasses opinion and argument, factoid and reason, as well as aspect and context (note that in the merged classes also the extraction and presentation of answers will be rather similar). The distribution of the resulting seven subclasses is shown in Table 3.

Classifiers for the comparative subclasses. To classify comparative questions into the fine-grained subclasses, we use neural BERT and CNN models since they performed best in pilot experiments (hyperparameters identical to Section 4). In particular, the BERT classifier is set up in a one-vs-rest manner (i.e., one model trained for every subclass) [5], while the CNN classifier is using a multi-label approach (lower computational effort at the same effectiveness as a one-vs-rest CNN). Instead of a softmax activation, a sigmoid activation function is used.

Table 3: Extended set of comparative Yandex questions with merged subclasses (6,250 questions in total).

Opinion/argument	4,101 (66%)	Preference	4,351 (70%)
Reason/factoid	2,074 (33%)	Direct	3,511 (56%)
Context/aspect	1,675 (27%)	Superlative	484 (8%)
Method	332 (5%)		

As the seven fine-grained classes are not mutually exclusive (i.e., questions have several labels) and since there is no intricate result page format change at stake (whether a question is comparative will be classified before), we do not treat the fine-grained classes as a precision-oriented classification. Instead, we optimize the hyperparameters for the micro-averaged F1—the common practice for multi-label classification [23, 44].[5]

The precision-recall curves of the classifiers on the 5,000 questions training set are shown in Figure 2. The largest subclasses of the interesting and probably also challenging to answer non-factoid comparative questions (i.e., opinion or argument) can be identified very reliably, as well as whether a preference was stated and whether comparison items are mentioned directly. In contrast, with relatively way fewer available examples, classifying whether aspects or context are mentioned seems to be hard with the BERT-based models being slightly better than the CNN models on some subclasses (but the CNN models require less computational effort).

6 EXPERIMENTS AND FINDINGS

We report the results of the different classification models described in Sections 4 and 5 on the respective test sets and then also conduct a qualitative analysis of the comparative questions in the year-long Yandex query log. This provides first insights into what comparative information needs users try to solve with a search engine.

Identifying comparative questions. Table 4 reports the classification results on the test set (10,000 questions). Not surprisingly, the individual models lose from 1% (logistic regression) up to 5% (BERT) in the recall, as well as the ensembles—from 3% (Ensemble-B+L) up to 5% (Ensemble-B+C+L).

[5]CNN: 200 filters, 4 epochs; all other parameters identical to Section 4.

Table 4: Classification results on the test set. Goal is the best recall at a precision of 1.0 on the comparative class (decision threshold in brackets). All classifiers achieve at least 0.98 precision and 1.0 recall for the non-comparative class.

Individ. model	Recall	F1	Ensembles	Recall	F1
Logistic (0.418)	0.54	0.70	Ens.-B+L (0.632)	0.60	0.75
CNN (0.99447)	0.52	0.68	Ens.-C+L (0.418)	0.59	0.74
BERT (0.99766)	0.44	0.61	Ens.-B+C+L (0.99447)	0.55	0.71
Rules (R1–7)	0.44	0.61			

We then test the relatively faster Ensemble-C+L "in the wild" by classifying the 1.5 billion questions from the Yandex log and manually check the assigned labels for another 5,000 comparative questions: about 1% misclassifications (cf. Section 5).

Fine-grained classification of comparative questions. Table 5 reports the classification results of the neural models on the test set for the fine-grained comparative subclasses. The respective models are trained in the settings and with the hyperparameters as described in Section 5. The three prevalent subclasses in the training dataset, *opinion/argument*, *preference*, and *direct* (each more than 3,500 training samples) achieve the best classification results. An exception is the underrepresented *superlative* subclass, which is identified relatively well by the models, probably due to the presence of adjectives and adverbs in a superlative form that are captured by the POS tag feature type.

Volumes, dynamics, and topics. To analyze monthly distributions and seasonal effects of the comparative questions in the year-long Yandex log, we apply the (almost) "perfect precision" Ensemble-C+L on the filtered 1.5 billion questions. Figure 3 shows the numbers of comparative questions per month as identified by the classifier (dark shade) along with an estimated total (light shade), based on the classifier's recall of 0.59 on the test set. The estimated ratio of the comparative questions is close to 3% throughout the year (we obtained 2.8% by random sampling) and shows an upward trend within the volume of all questions submitted to Yandex.

The comparative questions most frequently submitted to Yandex are shown in Table 6. A substantial part of them have the form `Which <item> is better/best to buy/choose/watch?` (many recalled by rule (R1)). Such questions target an informed choice, calling for opinions and arguments as pros and cons; they are hard to answer since they do not directly specify the items to be compared but require an analysis of all possible options within a set of items. Also, these questions often do not specify a comparison aspect and hence require to consider all involved items' features.

Table 5: Results of the comparative subclass classifications on the test set of the Yandex comparative questions.

	CNN			BERT		
Subclass	Prec.	Rec.	F1	Prec.	Rec.	F1
Opinion/argument	0.93	0.90	0.92	0.92	0.91	0.91
Reason/factoid	0.85	0.79	0.82	0.82	0.89	0.86
Context/aspect	0.88	0.52	0.62	0.75	0.74	0.74
Method	0.79	0.80	0.79	0.75	0.82	0.78
Preference	0.97	0.98	0.97	0.96	1.00	0.97
Direct	0.95	0.96	0.96	0.95	0.98	0.97
Superlative	0.93	0.79	0.86	0.92	0.86	0.89
Micro average	0.92	0.88	0.90	0.90	0.93	0.91

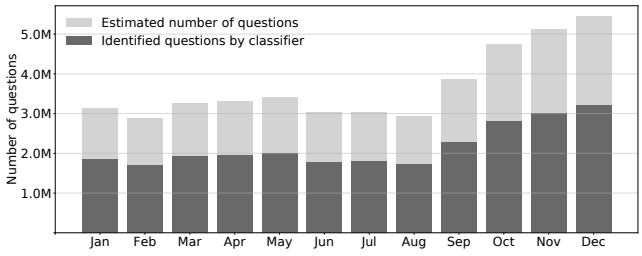

Figure 3: Monthly distribution of the comparative questions in the Yandex log.

To gain further insights, we categorize the recalled comparative Yandex questions into a scheme resembling the categories used on the Otvety platform. Following Völske et al. [40], we use a multinomial Naïve Bayes classifier trained on the Otvety data (14 merged topical categories), to categorize the comparative questions in the Yandex log. The categories with the relatively most comparative questions (ratio to the overall amount of questions in the category) are consumer electronics, followed by cars & transportation, home & garden, and education (cf. Table 7). As Figure 4 with the absolute numbers shows, the consumer electronics category exhibits the

Table 6: The ten most frequently asked comparative questions in the Yandex log.

Comparative question query	# Occur.
`Which pilot was the first to surpass a supersonic speed?`	176,372
`Which comedy is it better/best to watch?`	39,039
`Which is better, Xbox or PS?`	26,781
`Which tablet is it better/best to buy?`	24,443
`Anti-radar, which one is better?`	21,483
`Which phone is it better/best to buy?`	20,550
`Which antivirus is better/best?`	19,634
`What is the difference between a netbook and a laptop?`	18,165
`Which British colony was latest to receive independence?`	17,274
`Which laptop is it better/best to buy?`	16,775

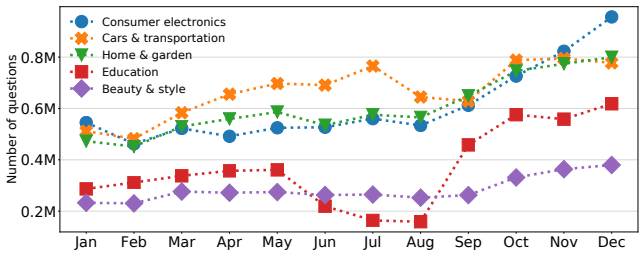

Figure 4: Yearly trend of the number of comparative Yandex questions in the largest topical categories.

Table 7: Total number of questions in millions, percentage of comparative questions, and the most frequently asked comparative questions per topical category in the Yandex log.

Question category	Quest. mln.	Comp. %	Most frequently asked question
Consum. electronics	105.4	6.3	Which tablet is it better/best to buy?
Cars & transport.	143.7	5.2	Anti-radar, which one is better?
Home & garden	166.7	4.0	Which vacuum cleaner is it best to buy?
Education	101.8	3.9	Which pilot was the first to surpass a supersonic speed?
Beauty & style	93.7	3.3	When is it best to cut hair?
Sports	45.8	3.1	Which time of the day is most suitable for doing sports?
Family & relationsh.	68.5	2.7	What is the best way to commit suicide?
Health	128.2	2.4	When is it best to conceive a baby?
Adult	53.9	2.3	What is the difference between men and women friendships?
Business & finance	133.7	2.0	In which bank is it best to take a loan?
Computers & intern.	152.4	2.0	Which antivirus is the best?
Society & culture	95.1	1.8	Which British colony was latest to receive independence?
Entertain. & music	90.2	1.5	Which comedy is it best to watch?
Games & recreation	122.0	1.3	Which is better, Xbox & PS?

largest increase at the end of the year—the number of comparative questions submitted in December doubles the February number. This indicates a clear seasonal trend: people tend to purchase electronics closer to the Christmas and New Year's holidays. The Russian school summer break from June through August explains the significant drop in education questions during these months, while in September and October they are asked almost as often as consumer electronics questions. Most of the topical categories remain constant or undergo a decrease during the summer months, indicating a stagnation or drop in online activities during holidays.

To dig a bit deeper into seasonal patterns, we also look at changes in the most frequent questions through the year. In March, the question When is it better to jog, in the mornings or in the evenings? is not among the top 20 of the most frequently asked ones, but it jumps to rank 13 in April (probably because of the more comfortable weather conditions and the approaching summer bathing season), stays at rank 11 in May, and disappears in June. Similarly, the question Which camera is it better/best to buy? is the 13th most frequent question in June, then moves down to rank 17 in July, and stays at rank 18 in August. The question What place at the Black Sea is better/best to go for vacation? reaches rank 8 in May, moves up to rank 3 in June, goes down to rank 6 in July, and leaves the top 20 in August, coinciding with the summer vacations. The mushroom picking season is indicated by the question How can one distinguish honey fungi from deadly skullcaps? jumping to rank 8 in September from out of the top-50 in August while the approaching winter is indicated by the question Which tires are better/best for winter? reaching rank 7 and Which is better, winter tire with metal studs or without? reaching rank 12 in October from out of the top-50 in September. Interestingly, the question Which pilot was the first to surpass a supersonic speed? was the most frequently asked question throughout the entire year, occupying rank 1 in every month except for January—an observation which we cannot really explain except that 2012 was the 65th anniversary of the achievement. The quite delicate questions of asking for the best ways of committing suicide (see Table 7) appears in January at rank 7, in March at rank 9, moves down to rank 12 in April, and disappears from the top-20 for the rest of the year. Questions of this kind should be identified by the search engine and treated with the appropriate care, however, this is out of the scope of this study.

Answering comparative questions. The above insights about the comparative questions' types and their categorical and temporal distribution can help a search engine to better understand the respective information needs and, in particular, to present the answers in an appropriate way. While pro/con answers to the most frequently asked questions could be cached, comparative questions also have a very long "tail" of rather rare intents. Our analysis of the compared items and the question categories shows that the comparison interests reach way beyond the traditionally studied areas of consumer products or factoid questions.

Our study of the comparative web search questions reveals that more than 65% of the questions are non-factoid (cf. Table 3) and demand argumentation and opinions in an answer (e.g., Which is better, Xbox or PS? or How dogs are better than cats?). One possible approach to tackle such questions is to extract "ready-to-use" answers from question answering platforms.

To test how good such an extraction approach might work, we index the cleaned set of all 5.5 million Otvety questions with a selected "best answers" with Elasticsearch's (BM25 as retrieval model). The 4,101 comparative Yandex questions labeled as opinion/argument are then used as search queries against this index (stop words removed) and human assessors labeled the answer to the top-ranked Otvety question as relevant or not for the Yandex question. It turns out that for about 48% of the comparative opinion/argument questions submitted to Yandex the top-ranked Otvety question with a best answer is relevant. In future work, it might thus be interesting to further investigate this possibility of extracting answers to non-factoid comparative questions from question answering platforms but also to further analyze the other half of the non-factoid questions that are probably not directly answerable using question answering platforms.

7 CONCLUSIONS

We have studied comparative questions submitted to Yandex over the period of one year. Such comparative questions form a non-negligible portion of the questions Yandex received (about 2.8%) and our study showed that the comparison intents reach far beyond just comparing products to buy or just expecting simple facts as answers (more than 65% of the comparative questions are clearly non-factoid). Only for about a half of the non-factoid questions, a good answer can be found on the Russian community question answering platform Otvety. Thus, if a search engine decided to support comparative questions in their entirety with direct answers, a focus on products or just relying on the engine's knowledge graph and online question answering platforms might not suffice.

Interesting directions for future work could be the development of approaches to automatically extract the compared items and the comparison aspects from comparative questions or the summarization / explanation of comparative answers for the non-factoid questions (e.g., by retrieving opinions or arguments on the web

that support a possible answer). This could then also improve comparative question handling in voice-only interfaces.

ACKNOWLEDGMENTS

This work has been partially supported by the DFG through the project "ACQuA: Answering Comparative Questions with Arguments" (grants BI 1544/7-1 and HA 5851/2-1) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999). We thank Yandex and Mail.Ru for granting access to the data. The study was partially conducted during Pavel Braslavski's research stay at the Bauhaus-Universität Weimar in 2018 supported by the DAAD. We also thank Ekaterina Shirshakova and Valentin Dittmar for their help in question annotation.

REFERENCES

[1] E. Agichtein, S. Cucerzan, and E. Brill. Analysis of Factoid Questions for Effective Relation Extraction. In *Proceedings of SIGIR 2005*. 567–568.
[2] M. Bendersky and W. Bruce Croft. Analysis of Long Queries in a Large Scale Search Log. In *Proceedings of the Workshop WSCD at WSDM 2009*. 8–14.
[3] P. Berezovsakaya and V. Hohaus. The Crosslinguistic Inventory of Phrasal Comparative Operators: Evidence from Russian. *Proceedings of FASL* 23, (2015), 1–19.
[4] P. Berezovskaya. Acquisition of Russian Comparison Constructions: Semantics Meets First Language Acquisition. In *Proceedings of ConSOLE 2013*. 45–65.
[5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer. 2006.
[6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching Word Vectors with Subword Information. *TACL* 5, 1 (2017), 135–146.
[7] J. Bresnan. Syntax of the Comparative Clause Construction in English. *Linguistic Inquiry* 4, 3 (1973), 275–343.
[8] J. Burger, C. Cardie, V. Chaudhri, R. Gaizauskas, S. Harabagiu, D. Israel, C. Jacquemin, Chin-Yew Lin, S. Maiorano, G. Miller, et al. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). In *Document Understanding Conferences Roadmapping Documents*. 2001. 1–35.
[9] Yu Cao, M. Fang, and D. Tao. BAG: Bi-Directional Attention Entity Graph Convolutional Network for Multi-Hop Reasoning Question Answering. In *Proceedings of NAACL 2019*. 357–362.
[10] L. B. Chilton and J. Teevan. Addressing People's Information Needs Directly in a Web Search Result Page. In *Proceedings of WWW 2011*. 27–36.
[11] R. Das, A. Godbole, D. Kavarthapu, Z. Gong, A. Singhal, Mo Yu, X. Guo, T. Gao, H. Zamani, M. Zaheer, and A. McCallum. Multi-Step Entity-Centric Information Retrieval for Multi-Hop Question Answering. In *Proceedings of MRQA at EMNLP 2019*. 113–118.
[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL 2019*. 4171–4186.
[13] Y. Feldman and R. El-Yaniv. Multi-Hop Paragraph Retrieval for Open-Domain Question Answering. In *Proceedings of ACL 2019*. 2296–2309.
[14] Y. Gal and Z. Ghahramani. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *Proceedings of NIPS 2016*. 1019–1027.
[15] I. Guy. Searching by Talking: Analysis of Voice Queries on Mobile Web Search. In *Proceedings of SIGIR 2016*. 35–44.
[16] S. Iyer, N. Dandekar, and K. Csernai. First Quora Dataset Release: Question Pairs. (2017). https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs.
[17] A. Jain and P. Pantel. Identifying Comparable Entities on the Web. In *Proceedings of CIKM 2009*. 1661–1664.
[18] N. Jindal and B. Liu. Identifying Comparative Sentences in Text Documents. In *Proceedings SIGIR 2006*. 244–251.
[19] N. Jindal and B. Liu. Mining Comparative Sentences and Relations. In *Proceedings of AAAI 2006*. 1331–1336.
[20] Y. Kim. Convolutional Neural Networks for Sentence Classification. In *Proceedings of EMNLP 2014*. 1746–1751.
[21] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of ICLR 2015*.
[22] O. Kolomiyets and M.-F. Moens. A Survey on Question Answering Technology from an Information Retrieval Perspective. *Information Sciences* 181, 24 (2011), 5412–5434.
[23] A. Kulkarni, N. R. Uppalapati, P. Singh, and G. Ramakrishnan. An Interactive Multi-Label Consensus Labeling Model for Multiple Labeler Judgments. In *Proceedings of AAAI 2018*. 1479–1486.
[24] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova, L. Jones, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, and S. Petrov. Natural Questions: A Benchmark for Question Answering Research. *TACL* (2019).
[25] T. W. Lauer and E. Peacock. An Analysis of Comparison Questions in the Context of Auditing. *Discourse Processes* 13, 3 (1990), 349–361.
[26] S. Li, C.-Y. Lin, Y.-In Song, and Z. Li. Comparable Entity Mining from Comparative Questions. In *Proceedings of ACL 2010*. 650–658.
[27] J. Mizuno, T. Akiba, A. Fujii, and K. Itou. Non-factoid Question Answering Experiments at NTCIR-6: Towards Answer Type Detection for Real World Questions. In *Proceedings of NTCIR 2007*.
[28] D. Moldovan, M. Paşca, S. Harabagiu, and M. Surdeanu. Performance Issues and Error Analysis in an Open-Domain Question Answering System. *TOIS* 21, 2 (2003), 133–154.
[29] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop CoCo at NIPS 2016*.
[30] Bo Pang and R. Kumar. Search in the Lost Sense of "Query": Question Formulation in Web Search Queries and its Temporal Changes. In *Proceedings of ACL 2011*. 135–140.
[31] C. Qu, L. Yang, W. Bruce Croft, F. Scholer, and Y. Zhang. Answer Interaction in Non-factoid Question Answering Systems. In *Proceedings of CHIIR 2019*. 249–253.
[32] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic Routing Between Capsules. In *Proceedings of NIPS 2017*. 3856–3866.
[33] M. Schildwächter, A. Bondarenko, J. Zenker, M. Hagen, C. Biemann, and A. Panchenko. Answering Comparative Questions: Better than Ten-Blue-Links?. In *Proceedings of CHIIR 2019*. 361–365.
[34] I. Segalovich. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. In *Proceedings of MLMTA 2003*. 273–280.
[35] S. Somasundaran, T. Wilson, J. Wiebe, and V. Stoyanov. QA with Attitude: Exploiting Opinion Type Analysis for Improving Question Answering in On-line Discussions and the News. In *Proceedings of ICWSM 2007*.
[36] M. Spenke, Ch. Beilken, and T. Berlage. FOCUS: The Interactive Table for Product Comparison and Selection. In *Proceedings of UIST 1996*. 41–50.
[37] S. Staab and U. Hahn. Comparatives in Context. In *Proceedings of AAAI 1997*. 616–621.
[38] A. von Stechow. Comparing Semantic Theories of Comparison. *Journal of Semantics* 3, 1-2 (1984), 1–77.
[39] Ji.-T. Sun, X. Wang, D. Shen, H.-J. Zeng, and Z. Chen. CWS: A Comparative Web Search System. In *Proceedings of WWW 2006*. 467–476.
[40] M. Völske, P. Braslavski, M. Hagen, G. Lezina, and B. Stein. What Users Ask a Search Engine: Analyzing One Billion Russian Question Queries. In *Proceedings of CIKM 2015*. 1571–1580.
[41] I. Weber, A. Ukkonen, and A. Gionis. Answers, not Links: Extracting Tips from Yahoo! Answers to Address How-To Web Queries. In *Proceedings of WSDM 2012*. 613–622.
[42] R. W. White, M. Richardson, and W.-tau Yih. Questions vs. Queries in Informational Search Tasks. In *Proceedings of WWW 2015*. 135–136.
[43] L. Xiao, H. Zhang, W. chen, Y. Wang, and Y. Jin. MCapsNet: Capsule Network for Text with Multi-Task Learning. In *Proceedings of EMNLP 2018*. 4565–4574.
[44] B. Yang, J.-T. Sun, T. Wang, and Z. Chen. Effective Multi-Label Active Learning for Text Classification. In *Proceedings of SIGKDD 2009*. 917–926.
[45] H. Yu and V. Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of EMNLP 2003*. 129–136.
[46] E. Yulianti, R.-Ch. Chen, F. Scholer, W. Bruce Croft, and M. Sanderson. Document Summarization for Answering Non-Factoid Queries. *TKDE* 30, 1 (2018), 15–28.