A "Suggested" Picture of Web Search in Turkish

ERDEM SARIGIL and OGUZ YILMAZ, Bilkent University
ISMAIL SENGOR ALTINGOVDE, Middle East Technical University
RIFAT OZCAN, Turgut Ozal University
ÖZGÜR ULUSOY, Bilkent University

Although query log analysis provides crucial insights about Web users' search interests, conducting such analyses is almost impossible for some languages, as large-scale and public query logs are quite scarce. In this study, we first survey the existing query collections in Turkish and discuss their limitations. Next, we adopt a novel strategy to obtain a set of Turkish queries using the query autocompletion services from the four major search engines and provide the first large-scale analysis of Web queries and their results in Turkish.

CCS Concepts: • **Information systems** → **Web searching and information discovery**; **Web log analysis**

Additional Key Words and Phrases: Turkish query characteristics, query spelling correction

ACM Reference Format:
Erdem Sarigil, Oguz Yilmaz, Ismail Sengor Altingovde, Rifat Ozcan, and Özgür Ulusoy. 2016. A "suggested" picture of Web search in Turkish. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 15, 4, Article 24 (May 2016), 11 pages.
DOI: http://dx.doi.org/10.1145/2891105

1. INTRODUCTION

Query log analysis reveals a lot of valuable information regarding the search interests and behavior of Web users that may be exploited for various purposes, such as query expansion, suggestion, and correction [Silvestri 2010]. Commercial search engines occasionally release query logs for supporting research in various directions; however, such datasets are strictly anonymized not to disclose neither the query text nor the clicked documents. Even on the rare occasions where such query logs include query strings (as in the case of AltaVista and Excite logs from previous decades [Silvestri 2010] and the AOL log from 2006 [Pass et al. 2006]), the logs are typically collected from English-speaking users and hence are mostly in English. For several other languages, such as Turkish, which we focus on in this study, large-scale and public query logs are quite scarce or simply nonexistent.

In this study, we first review the existing query collections in Turkish and show that they all include a small number of queries created within an experimental setup

This work was funded by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under grant 113E065.
Authors' addresses: E. Sarigil, O. Yilmaz, and Ö. Ulusoy, Dept. of Computer Engineering, Bilkent University, Ankara, Turkey; emails: {esarigil, oguzy, oulusoy}@cs.bilkent.edu.tr; I. S. Altingovde, Dept. of Computer Engineering, Middle East Technical University, Ankara, Turkey; email: altingovde@ceng.metu.edu.tr; R. Ozcan (corresponding author), Dept. of Computer Engineering, Turgut Ozal University, Ankara, Turkey; email: rozcan@turgutozal.edu.tr.

DOI: http://dx.doi.org/10.1145/2891105

(Section 2). Next, we adopt a novel strategy to obtain a large set of Turkish queries using the public query autocompletion services from four major search engines on the market used for research purposes: Bing, Google, Yahoo, and Yandex. These queries are further submitted to Google's Keyword Planner Tool, and the majority of them are found to have a nonzero search volume, providing clear evidence that they are real queries submitted by search engine users. Based on this unique dataset, our contributions are as follows:

—We first analyze to what extent the collected query suggestions can represent the search trends of Turkish-speaking Web users (Section 3). In particular, we assess the language specificity of the suggestions using both an automated technique based on an open-source Turkish NLP library and extensive user studies over thousands of queries.
—Second, we identify search trends of Turkish users in terms of the query intent, entity types mentioned in the query, and query category (Section 4).
—As a third contribution, we focus on the query results and report various analyses regarding the coverage of Turkish Web space in the query results and the evolution of query results within a 6-month observation period (Section 5).
—Finally, we investigate the state art of the art for addressing spelling errors specific to Turkish by submitting several queries including automatically injected errors to different search engines (Section 6).

Our work provides the first large-scale analysis of real user queries in Turkish. By doing so, we identify directions and areas for further research to improve the search experience in Turkish. Furthermore, our work may serve as a guideline for conducting similar studies for other languages where large-scale query logs are not publicly available.

2. A SURVEY OF TURKISH QUERY COLLECTIONS

To the best of our knowledge, there is no publicly available query log from major search companies that can be used for characterizing Turkish Web search trends. Thus, the researchers have compiled Turkish query sets via different ways for various research goals as follows:

—Bitirim et al. [2002] created a set of 17 query topics to analyze the retrieval performance of four local search engines. The queries have retrieved a total of 971 documents that were annotated for relevance. Furthermore, the top 1,000 results were retrieved for the top 5 most popular search terms to compare the novelty, coverage, and recency of the search engines.
—In Demirci et al. [2007], 12 single-term queries were defined, and their top 20 results were retrieved and annotated to evaluate the performance of five international search engines.
—To create a set of 24 queries, Tokgoz et al. [2013] first determined eight categories from six popular Turkish news portals and then chose three queries per category. These queries were submitted to four major search engines (Bing, Google, Yahoo, and Yandex) for four consecutive weeks, and the top 20 results were retrieved separately for Web and image search options. The resulting 7,670 and 7,680 links from Web and image search, respectively, were used to evaluate the retrieval performance in terms of the precision.
—Turker and Bitirim [2015] compiled a set of 60 queries that include 30 frequently misspelled terms from the Web and their correct forms. These queries were submitted to Bing, Google, and Yahoo, and the top 20 results were retrieved. The resulting 3,586 documents were used to evaluate the retrieval performance of these engines.

—In Can et al. [2008], a set of 72 query topics, along with relevance judgments, was created. The queries were formulated and evaluated by 33 native speakers following a TREC-like approach—that is, for each query, the top 100 results were retrieved using various ad hoc retrieval functions to create a document pool for the subsequent relevance annotation, which yielded 104.3 relevant documents on average.

Although all of these earlier studies provide valuable insights, the query sets typically include just a few dozen queries that are either created by the authors themselves or by participants in an experimental setup. In contrast, our query set involves a large number of real queries from search engine users (as verified by the search volume analysis presented in Section 3), which means that we construct, analyze, and share with the research community the largest and most realistic Turkish Web query collection to date. This unique dataset also enables us to focus on the properties of Turkish query strings for the first time in the literature.

3. PREFIX-BASED QUERY SUGGESTION COLLECTION

Autocompleting the query string that is being typed is a popular feature of commercial search engines to assist users. Such autocompletions (suggestions) are further utilized for various novel tasks in recent studies. For instance, in Chelaru et al. [2013], a large dataset obtained from a major search engine’s query autocompletion service is exploited for investigating sentiment in queries for controversial topics. Here, we leverage autocompletion services to compile a representative set of queries from Turkish-speaking Web users. In what follows, we first describe our data collection methodology from the four major search engines. Next, we present automatic and manual results for identifying the queries within the Turkish search space—that is, those likely to be submitted by Turkish-speaking Web users.

As our preliminary experiments revealed that Yahoo suggestions include a large number of non-Turkish queries, we applied a two-stage strategy where we first collected suggestions from Google, Yandex, and Bing, and afterward from Yahoo, by exploiting the terms encountered in the queries from the former three search engines. For all search engines, we collected query suggestions (instant autocompletions) using their Turkish front-end (i.e., the search page with the *tr* country extension) and/or setting the language preference as Turkish. We collected only full query suggestions, not partial suggestions that occur for the last term being typed, which sometimes happens when the query prefix does not match any of the queries in the suggestion database. Note that we treat Yahoo and Bing separately. although Yahoo uses the organic results from Bing, as at the time of conducting these experiments they yielded reasonably different query suggestions and retrieval results.

Query suggestion collection from Google, Yandex, and Bing. We first collected the top 10 query suggestions for all queries that matched the prefix *(letter)(letter)(letter)*—that is, all legitimate 3-letter combinations in Turkish. Additionally, we also considered queries with the prefix *(letter)(blank)* and *(letter)(letter)(blank)* (i.e., the query “o ses turkiye” matches the former and “as tv” matches the latter pattern). Given that the Turkish alphabet includes 29 letters, these 3-letter combinations and the additional patterns with blanks yielded 25,172 prefixes to be submitted to each search engine. We used no more than 3-letter combinations to avoid overloading the search engines.

We noticed that some prefixes yielded fewer than 10 suggestions, which implies that there is no point in further refining these seeds. In contrast, for the prefixes yielding exactly 10 suggestions, there can be a wealth of queries that we are missing. To address this issue without overloading the search engines, we decided to submit longer prefixes for only word stems that are known (and relatively popular) in Turkish. We first created a term list from all unique words appearing in the suggestions described earlier. We

merged this list with an additional word list that involves the unique terms extracted from a news dataset including 408,305 articles published between 2001 and 2005 in a popular Turkish newspaper, *Milliyet* [Can et al. 2008]. We created 4-letter and 5-letter stems from these terms (we did not consider more than 5 letters, as a previous work reported that the first 5-letter stems are quite effective for Turkish IR [Can et al. 2008], and thus they can also be representative enough to collect queries). We also considered combinations like "3-letter full words + blank" and "4-letter full words + blank." This procedure yielded an additional 112,454 combinations to send to search engines. In the end, we sent all 137,626 (25,172+112,454) combinations to Google, Yandex, and Bing, and obtained 686,361, 193,174, and 525,277 unique query suggestions, respectively.

Query suggestion collection from Yahoo. We first attempted to use exactly the same strategy described earlier to obtain the suggestions from Yahoo. However, it turned out that for 3- or even 4-letter prefixes, a significant amount of the suggested queries were non-Turkish. Therefore, we decided to provide further clues for this search engine and compiled a list of all *single-term query suggestions* from Google, Yandex, and Bing datasets. We ended up with 185,439 terms that are most likely to be in Turkish queries and submitted these as prefixes to Yahoo. This process yielded 984,529 queries.

Although Yahoo seems to provide a large number of queries, our manual inspection (more than thousands of queries) revealed that a very high percentage of these queries are still in English (or, more generally, non-Turkish). In the following, we discuss a methodological categorization of the queries as Turkish or not and the results of automatic and manual assessments in this respect.

Identifying query suggestions for Turkish users. We categorized each query suggestion into one of the following classes:

—*Purely/partially Turkish queries*: Queries with at least one term in the Turkish language (say, in the dictionary/thesaurus), lists of Turkish proper names, or well-known Turkish Web site or brand names. For example, "araba oyunlari" (*car games*) is a purely Turkish query, of which both terms (stems) are in the dictionary. Additionally, queries with popular Web site names like "oyna65.com" (*play65.com*), "ensonhaber.com" (*latestnews.com*), and "daybuyday.com" are also considered Turkish, with the latter being the name of a popular Turkish shopping site with a fully English name (day-buy-day) but interests Turkish speaking users in the first place. Finally, queries with at least one Turkish term, such as "youtube mp3 dönüstürücü" (*youtube mp3 converter*), "fiat doblo fiyatlari" (fiat doblo *prices*) and "proshow gold indir" (proshow gold *download*) also fall into this category. In such queries, usually there is one popular global entity in a non-Turkish language (e.g., YouTube, Fiat Doblo, and Proshow Gold) along with some Turkish terms.
—*Non-Turkish queries with global entities*: In this case, the query only involves a globally popular entity (e.g., "fiat doblo"). We think that even though these queries do not involve a Turkish term, they can be asked by Turkish users and be answered with pages that are in Turkish Web space. Indeed, the existence of other query suggestions with additional terms (e.g., fiat doblo *prices*) provides evidence in this direction—that is, such global entities can be directly queried by Turkish-speaking users.
—*Purely non-Turkish queries*: All queries that do not fall into the preceding categories are considered non-Turkish. Of course, some of these queries may still have been submitted by Turkish users; however, without having actual search logs, it is impossible to be sure whether these queries can represent the "Turkish search space" or whether they are generated by the search engine from a global suggestion database when the submitted query prefix does not match any previous local searches.

We employed an open-source NLP tool for Turkish—Zemberek (http://code.google.com/p/zemberek/)—to determine the queries that fall into the first category (i.e., those

Table I. Fraction of the Purely/Partial Turkish Queries Based on the Zemberek NLP Package

	Google	Yandex	Yahoo	Bing
Total Number of Suggestions	686,361	193,174	984,529	525,277
Average Number of Suggestions per Prefix	5.0	1.4	5.3	4.9
Purely/Partially TR (%)	56%	76%	19%	82%

Table II. User Study Results for 21,930 Query Suggestions from Four Search Engines

	Google	Yandex	Yahoo	Bing
Number of Suggestions	6,438	4,445	5,066	5,981
Purely/Partially TR (%)	47.6%	60.2%	10.5%	65.90%
Non-TR with Global Entities (%)	8.5%	9.1%	15.7%	7.89%
Purely Non-TR (%)	43.9%	30.8%	73.8%	24.3%

including at least one Turkish term). Note that this tool cannot recognize terms (e.g., Turkish entity names) that are not in a typical thesaurus. In Table I, for each suggestion set, we provide the percentage of Turkish queries as identified by Zemberek. Clearly, Yahoo has the smallest fraction of queries classified as Turkish, whereas Bing yields the largest fraction of Turkish query suggestions.

For a more detailed analysis that takes into account the queries with Turkish proper names and the global entities, we conducted an extensive user study. In particular, we randomly sampled 800 query prefixes from the 185,439 terms that are used for seeding Yahoo and compiled a list that includes the shuffled set of the top 10 suggestions from all four search engines for each such prefix. The final list included 21,930 query suggestions. Five human assessors (computer science researchers who natively speak Turkish) were asked to categorize each query suggestion into one of the three query categories described earlier.

In Table II, we provide the percentage of queries for each category and each search engine, along with the total number of suggestions. Although the actual percentages vary with respect to Table I, the general trends are similar. It turns out that for Bing, Yandex, and Google, the percentages of Turkish queries are significantly larger than those for Yahoo, for which purely non-Turkish suggestions add up to more than 70%. It is also remarkable that Bing and Yandex suggest a higher percentage of purely partially TR queries than Google, although the latter search engine is claimed to have the largest market share in Turkey.

Query volume analysis. Since the methods for generating query suggestions are not disclosed by commercial search engines, it is crucial to justify that the query sets constructed using the preceding methodology include queries based on real user submissions and are not generated by other means (e.g., using the index content). To this end, following the practice in Chelaru et al. [2013], we submitted the queries employed in our user study, namely 21,930 queries, to Google’s Keyword Planner Tool (in batches of 1,000 queries). This tool is provided by Google to assist in choosing appropriate ad words for advertisement campaigns; besides other statistics, it reports the average monthly search volume (starting from November 2012 to date) of a given keyword query.

It turns out that 67% of the queries in our set have a nonzero average monthly search volume—that is, they are submitted at least 10 times or more by the users. Indeed, we observe that 7,423 queries (34% of the query set) are submitted more than 1,000 times per month on average. If we break up the volume information for each search engine, we find that 56%, 66%, 75%, and 76% of the queries suggested by Bing, Google, Yahoo, and Yandex have nonzero search volume, respectively. Note that some queries submitted to one of the other search engines may have never been issued to

Table III. Distribution of Queries Across 14 Yahoo Categories and the Category "Other"

Category	%	Category	%	Category	%
Entertainment	19.9	Education	3.5	News & Media	1.9
Business & Economy	12.5	Society & Culture	3.5	Government	1.8
Computer & Internet	12.0	Recreation & Sports	3.3	Social Science	1.7
Regional	6.7	Science	2.4	Reference	1.3
Health	3.7	Arts & Humanities	2.0	Other	23.8

Google, explaining why they have no volume information. Furthermore, we realized that some of these queries were time sensitive and hence might disappear by the time of our volume analysis (early 2014), which had been conducted much later than the construction of our query sets (early 2012). Despite the latter limitations, our analysis still verifies that the majority of our queries are submitted by real users and hence can be used as a basis for the analysis presented in this article. We provide the query set and user annotations at http://www.ceng.metu.edu.tr/~altingovde/trIR.htm.

4. CHARACTERISTICS OF TURKISH WEB QUERIES

In this section, we focus on the first two query categories described previously, namely purely/partially TR queries and non-TR queries with global entities, as they are most likely to represent the search trends of Turkish-speaking Web users. To this end, we randomly sampled 500 queries that are in one of these categories and are suggested by all four search engines. Furthermore, we sampled 125 queries that are suggested exclusively by each one of the four search engines. In the end, we created a set of 1,000 queries. We conducted a user study and asked four human judges to annotate three core aspects in this query set—entity type in the query, query intention and query category—as discussed in detail next:

—*Query entities*: We took into account four major types of entities: person, product, organization, and location. We also asked annotators to mark the cases when there were terms in addition to an entity itself (e.g., "Michael Jackson mp3s").
—*Query intention*: Following the practice in the literature, we classified queries into three types: informational, navigational, and transactional. As popular query suggestions tend to include a high percentage of entities, we provided annotators with a detailed set of rules for deciding on the intention of the queries with some entities as follows. Queries with a person entity (possibly accompanied by a name of a movie, song, etc.) were considered as informational (e.g., "kemal sunal" (Turkish actor) and "kemal sunal milyoner," where "milyoner" is the name of a movie). Queries with an organization entity, which could also be accompanied with location information and/or terms like "harita" (*map*), were navigational (e.g., "thy" (*Turkish airlines*) and "alkoclar otomotiv ankara" (alkoclar *automotive* ankara)). Queries only with the name of a product, including a movie, song, or TV series, were considered informational (e.g., "dövüs klübü" (*fight club*), "iphone 5," and "world of warcraft"). However, if they were accompanied with verbs like "al" (*buy*), "sat" (*sell*), "izle" (*watch*), "oyna" (*play*), or "indir" (*download*), they were annotated as transactional queries. Finally, we considered queries including only place names as informational (e.g., "ali pasa hamami" (ali pasa *hammam*) and "catalhöyük" (a historical town in Turkey)), whereas they were again classified as navigational when they were accompanied with terms like "harita" (*map*) (as in this case user usually looks for a specific map service in mind, i.e., Google Maps or Yandex Maps).
—*Query category*: We classified queries into one of the 14 top-level categories obtained from the Yahoo directory service (as shown in Table III). Queries that did not fit into any of these categories were labeled as "Other."

We find that a high fraction of queries, namely 64%, included at least one type of entity. This observation conforms to previous findings in the literature, as Guo et al. [2009] also report that 70% of the queries from a set of 1,000 manually inspected queries (obtained from a proprietary search log) included a named entity. Note that our query set may exhibit characteristics that are more similar to those of the head and torso queries encountered in a typical search log, as query suggestions are usually generated upon past popular queries. Still, given that such head and torso queries constitute half of the entire query volume that hits a search engine, we believe that our findings in this article strongly represent Turkish Web users' search trends.

The most popular entities in queries are from the product (20.1%) and person (19%) types, which include Turkish celebrities like “kemal sunal” (actor), “ece erken” (vj), “volkan demirel” (football player), and various products in the broadest sense (i.e., including software, games, movies, songs, electronic devices, and medicines). Organization-type entities are also quite popular (17.3%) in queries, whereas the popularity of location-type entities is almost one third of the other entity types (i.e., 7.6%).

Next, we discuss query intent analysis results. It turns out that the majority of Turkish queries, namely 67%, are informative, whereas 20% of them are navigational and the remaining 13% are transactional queries. Note that this analysis excludes adult/porn queries, as query suggestion services filter them. In one of the first papers that introduced the preceding query intent taxonomy, Broder [2002] reports the results from a manual analysis of 1,000 queries sampled from a query log and cleaned to remove non-English and sexually oriented queries. According to this study, informative, navigational, and transactional queries correspond to 48%, 20%, and 30% of the query sample, respectively. In another study, Rose and Levinson [2004] report that around 62% of the queries were informational, 13% navigational, and 24% transactional (called *resource* in their taxonomy). In comparison to these previous findings, Turkish queries seem to have slightly smaller fraction of transactional queries. This might be due to the bandwidth limitations (to download large files) or e-commerce practices of the Turkish users.

Finally, Table III shows the distribution of queries across 14 Yahoo categories. Around 23% of queries cannot be classified to any of these categories and thus are shown with the label “Other.” The dominant query category is Entertainment, covering 20% of the queries. The other two most popular categories are Business & Economy and Computer & Internet, taking shares of 13% and 12%, respectively. A similar analysis over the AOL log is reported in Beitzel et al. [2007], where it is found that 10% of queries fall into the Porn category, whereas 13% are in Entertainment. When porn queries are filtered, the latter value for Entertainment queries is computed as 14.4%, which is not so far from our finding. In this latter work, the Shopping category is also 13% (14.4% after removing Porn) and the Computer and Games categories sum up to 14% (15.5% after recomputation) of the queries, which again are similar figures to ours. Of course, it is hard to estimate the volume of the adult-related queries in our case; however, we can still conclude that the nonporn query streams in Turkish and English have certain similarities in terms of the query category distribution.

5. CHARACTERISTICS OF TURKISH QUERY RESULTS

In this section, we focus on the properties of the top 10 query results retrieved for the 1,000 queries analyzed in the previous section. To this end, we submitted these 1,000 queries to the four search engines and stored their top 10 result pages at three different points in time: February, May, and August 2012. We sought answers to the following questions: What is the overall number of results reported for Turkish queries? What is the fraction of results from sites with Turkish country extension (*tr*)? What is the fraction of results that have Turkish content? And finally, how do these findings evolve within time?

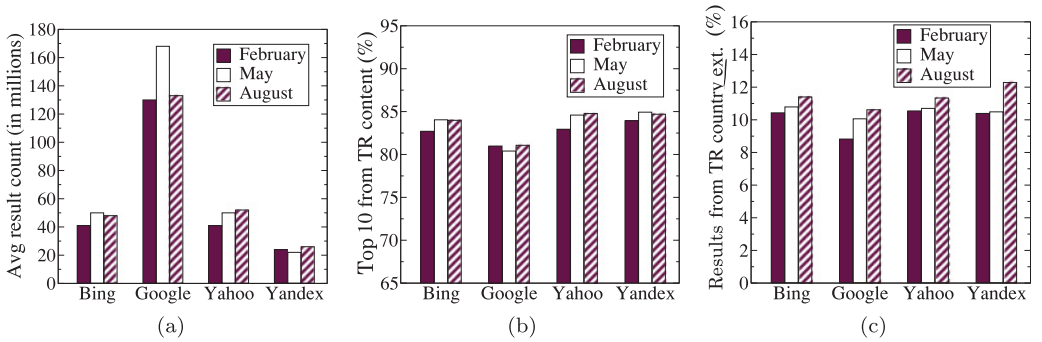

Fig. 1. Query result trends in time: average number of results (a), fraction of results with snippets in Turkish (b), and fraction of results (c) from TR domain.

In Figure 1(a), we show the average number of results per search engine and at each time point. It turns out that Google reports the largest number of results for Turkish queries and almost triples that of other search engines. Bing and Yahoo, not surprisingly, find a similar number of results, and Yandex, being the youngest search engine in the Turkish search market, reports almost half of the numbers reported by the former two competitors. It also seems that the number of results found for these queries slightly increases for Bing and Yahoo, and it is rather stable for Yandex. For Google, there is a big increase in May, but this seems rather temporary, and the average number of results drops in August (but to a point still higher than that in February). We speculate that Google might have removed some spam pages (or other similar pages) between these two points in time.

As the number of results reported by search engines usually serves as a rough approximation, we further analyzed the content of the top 10 query results. To this end, we used Zemberek to decide whether the snippet sentences retrieved per query result were in Turkish or not. A snippet was labeled as Turkish content if it included at least three Turkish terms recognized by Zemberek. We found that the majority of the search results (i.e., more than 80%) were pages in Turkish (Figure 1(b)). Although the percentages of Turkish results were similar for all search engines, this time Google was slightly behind all others, and Yandex seemed to have the largest fraction of results in Turkish. Interestingly, although Yahoo was found to be less effective in suggesting Turkish queries, it could still retrieve a comparable percentage of results from Turkish Web space with respect to its competitors. The Turkish content slightly increased for all search engines but Google within our observation period.

Finally, we investigatd the percentage of the top 10 query results that were from Web domains with the country extension *tr*. Figure 1(c) reveals that only a small fraction (around 10%) of query result URLs includes the *tr* extension. Given the results of previous experiment on the snippets, we conclude that most Turkish Web sites retrieved in the results prefer to have global domain names, but not those ending with the Turkish country extension.

6. SPELLING CORRECTION PERFORMANCE

Spelling correction is an important feature of a search engine, as up to 15% of submitted queries contain spelling errors [Croft et al. 2009]. To the best of our knowledge, no previous work has investigated the state-of-the-art performance of spelling correction for Turkish queries; however, such an analysis can provide important insight for improving the performance of spelling correction approaches for Turkish.

Table IV. Five Different Error Types Injected into the Queries

Error Type	Definition
Transposition	A randomly chosen character is swapped with the next character (e.g., acm → cam)
Substitution	A character is replaced by one of its neighbors on the keyboard (e.g., acm → avm)
Insertion	A neighbor of a randomly chosen character is inserted (e.g., acm → acvm)
Deletion	A randomly chosen character is deleted (e.g., acm → ac)
Diacritics	All Turkish characters (ğ, ü, ı, ş, ö, ç) are written without diacritics (e.g., özgür → ozgur)

Table V. Response of Search Engines to Queries in Typo-I, -II, and -III Sets

	Typo-I set					**Typo-II set**					**Typo-III set**				
Search Engine	AC	P0	P1	P2	P3	AC	P0	P1	P2	P3	AC	P0	P1	P2	P3
Google	468	83	93	352	4	304	61	97	512	26	219	93	82	522	84
Yandex	401	39	0	547	13	259	34	0	645	62	193	48	0	578	181
Yahoo!	259	302	12	402	25	150	198	2	450	200	98	162	6	414	320
Bing	434	219	9	314	24	328	156	2	403	111	252	138	7	348	255

To this end, we first constructed a query set containing 1,000 Turkish queries by manually choosing from the query suggestion collections described in Section 3. This set contained 500 queries that were suggested by all four search engines. Additionally, four different sets of 125 queries were selected from suggestions of each search engine exclusively. For the experiments in this section, we need to consider queries of which all terms are decided to be in Turkish (manually); therefore, we cannot use the query set employed in Sections 4 and 5.

Based upon this initial query set, we created our misspelled query sets by injecting five major types of errors, which are shown in Table IV. Note that the first four types of errors are known to correspond to more than 80% of all human misspellings [Croft et al. 2009]. We constructed three misspelled query logs—Typo-I, Typo-II, and Typo-III—as follows. For a query put into the Typo-I set, we randomly determined a position in the original (correct) query string and then randomly chose and applied one of the error types described in Table IV. The queries in the Typo-II and Typo-III sets included two and three typos that were injected in the same manner, respectively. Since our aim was to investigate the state-of-the-art spelling correction performance for Turkish queries with different levels of spelling errors, and search engines are likely to employ the most advanced software for this purpose, we submitted queries in each of these sets to four major search engines.

As we discussed in Section 2, query autocompletion is a popular search engine functionality not only for saving user from the burden of typing a query but also for preventing typos. For the cases where a corrected query cannot be suggested, or the user does not click on any suggestion but insists on submitting what she typed, Altingovde et al. [2012] observed four different types of actions (referred to as behavior patterns) that can be taken by the search engines. In pattern 0, the search engine does not suggest any alternative query and returns results using the original submitted query. If the search engine returns the results for an original query but suggests an alternative query (e.g., using the well-known *did you mean* phrase), then this is classified as pattern 1. In pattern 2, the search engine presents the results of an alternative query instead of the original query. In this case, a message is displayed that includes a link to the results of the original query (e.g., in Google, this message is like *showing results for <suggested query>, search instead for <original query>*). Pattern 3 denotes the case when the search engine cannot return any results.

We present the results for Typo-I, Typo-II, and Typo-III sets in Table V. The AC column denotes the number of misspelled queries for which the correct versions are suggested by an autocompletion mechanism while the query is typed by the user.

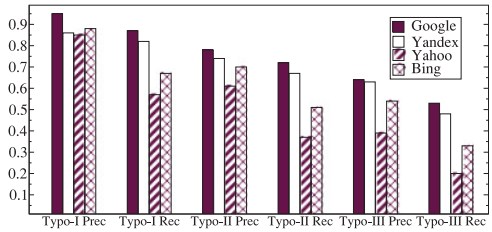

Fig. 2. Suggestion precision and recall for Typo-I, Typo-II, and Typo-III sets.

The remaining queries are associated with one of the patterns (P0 to P3) described earlier. From the results, we first observe that as the level of misspelling is increased, the number of queries that cannot be answered (pattern 3) increases considerably, especially for Yahoo and Bing. Overall, Google has the lowest number of queries with no results, and Yandex's performance is between Google and Yahoo/Bing. Especially for Typo-II and Typo-III sets, Google and Yandex provide corrections (via patterns 1 and 2) for similar numbers of queries. Note that all three search engines (Google, Bing, and Yandex) resolve more than 40% of the misspelled queries in the Typo-1 set via the autocompletion mechanism.

As a final analysis, we evaluated the quality of the suggested corrections provided via the autocompletion mechanism and patterns 1 and 2. To this end, we used suggestion precision and suggestion recall metrics. Suggestion precision is the ratio of the number of successfully spell-corrected queries to the number of queries with suggestions. Suggestion recall is defined as the ratio of the number of successfully spell-corrected queries to the number of all queries with typos.

Figure 2 shows the plot for suggestion precision and recall figures for Typo-I, Typo-II, and Typo-III sets. It is observed that as the level of spelling error is increased, precision and recall values decrease for all four search engines. Among them, Google and Yandex correct the errors with considerably higher precision than Yahoo and Bing. This finding can be expected: the latter two search engines are more active in Turkish market and hence are more likely to deploy customized tools/services for Turkish.

7. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

We reviewed the existing query collections in Turkish and found out that all include a small number of queries created within an experimental setup. As a remedy, we leveraged four major search engines' query suggestion services to conduct the first large-scale analysis of Turkish Web search queries. Our analysis sheds light on several questions: What percentage of the suggested queries can represent search interests of Turkish-speaking Web users? What are the characteristics of these queries? To what extent does Turkish content appear in the results of these queries? What is the state-of-the-art performance for resolving spelling errors in queries that involve only Turkish words? We address these questions by using automated methods and extensive user studies, in which several thousands of queries are examined.

Overall, the number of suggestions collected from four international search engines that are currently available in the Turkish search market is more than 2.3 million queries, which makes our dataset a precious resource to foster research in the Turkish IR field. We believe that our work in this article points to several interesting research directions. First, the user study conducted in Section 3 shows that it is not a trivial task to identify queries that are related to users from a certain region/country. Multilingual strategies that can take into account users' local search trends and blend these with the global trends in intelligent ways can be investigated. Second, the findings in Section 4

regarding certain aspects of Turkish Web queries can fuel further research in various areas. For instance, the high fraction of queries with entities implies that focusing on the techniques for entity/relationship extraction in Turkish can improve retrieval effectiveness. Finally, our findings in Section 6 reveal that current search engines are vulnerable when more than one spelling error occurs in Turkish queries. This might be due to the relatively smaller volume of Turkish queries available for training spelling correction systems for search engines. Thus, techniques that can employ other (external) resources (e.g., the translation of popular queries in other languages) can also be adapted for Turkish to improve spelling correction performance.

REFERENCES

Ismail Sengor Altingovde, Roi Blanco, Berkant Barla Cambazoglu, Rifat Ozcan, Erdem Sarigil, and Özgür Ulusoy. 2012. Characterizing Web search queries that match very few or no results. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*. 2000–2004.

Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, Ophir Frieder, and David A. Grossman. 2007. Temporal analysis of a very large topically categorized Web query log. *Journal of the Association for Information Science and Technology* 58, 2, 166–178.

Yiltan Bitirim, Yasar Tonta, and Hayri Sever. 2002. Information retrieval effectiveness of Turkish search engines. In *Proceedings of the 2nd International Conference on Advances in Information Systems*. 93–103.

Andrei Broder. 2002. A taxonomy of Web search. *SIGIR Forum* 36, 2, 3–10.

Fazli Can, Seyit Kocberber, Erman Balcik, Cihan Kaynak, Huseyin Cagdas Ocalan, and Onur M. Vursavas. 2008. Information retrieval on turkish texts. *Journal of the Association for Information Science and Technology* 59, 3 (2008), 407–421.

Sergiu Chelaru, Ismail Sengor Altingovde, Stefan Siersdorfer, and Wolfgang Nejdl. 2013. Analyzing, detecting, and exploiting sentiment in Web queries. *ACM Transactions on the Web* 8, 1, 6.

W. Bruce Croft, Donald Metzler, and Trevor Strohman. 2009. *Search Engines: Information Retrieval in Practice*. Pearson Education.

Rabia Gulcin Demirci, Vildan Kismir, and Yiltan Bitirim. 2007. An evaluation of popular search engines on finding Turkish documents. In *Proceedings of the International Conference on Internet and Web Applications and Services*. 61.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd Annual International SIGIR Conference on Research and Development in Information Retrieval*. 267–274.

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems*.

Daniel E. Rose and Danny Levinson. 2004. Understanding user goals in Web search. In *Proceedings of the 13th International Conference on World Wide Web*. 13–19.

Fabrizio Silvestri. 2010. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval* 4, 1–2, 1–174.

B. Tokgoz, Z. N. Ozcilnak, C. Cinar, M. T. Yalun, and Y. Bitirim. 2013. An evaluation of Turkish retrieval performance of popular search engines for Internet and image search by using common lists. In *Proceedings of the Conference on Digital Information and Communication Technology and Its Applications*. 148–153.

Seyda Turker and Yiltan Bitirim. 2015. A study on the effect of using the wrongly spelled and/or pronounced Turkish words on Seb search engines. *International Journal of Scientific Knowledge* 5, 8, 1–7.

Received April 2015; revised November 2015; accepted February 2016