Information Processing and Management 56 (2019) 228–246

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Improving educational web search for question-like queries through subject classification

Tolga Yilmaz[*,a], Rifat Ozcan[b], Ismail Sengor Altingovde[c], Özgür Ulusoy[a]

[a] Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey
[b] Ntent, Inc., Av. Diagonal 210, Barcelona 08018, Spain
[c] Department of Computer Engineering, Middle East Technical University, Ankara 06800, Turkey

ARTICLE INFO

Keywords:
Educational web search
Question classification
Search engine result page ranking
K-12

ABSTRACT

Students use general web search engines as their primary source of research while trying to find answers to school-related questions. Although search engines are highly relevant for the general population, they may return results that are out of educational context. Another rising trend; social community question answering websites are the second choice for students who try to get answers from other peers online. We attempt discovering possible improvements in educational search by leveraging both of these information sources. For this purpose, we first implement a classifier for educational questions. This classifier is built by an ensemble method that employs several regular learning algorithms and retrieval based approaches that utilize external resources. We also build a query expander to facilitate classification. We further improve the classification using search engine results and obtain 83.5% accuracy. Although our work is entirely based on the Turkish language, the features could easily be mapped to other languages as well. In order to find out whether search engine ranking can be improved in the education domain using the classification model, we collect and label a set of query results retrieved from a general web search engine. We propose five ad-hoc methods to improve search ranking based on the idea that the query-document category relation is an indicator of relevance. We evaluate these methods for overall performance, varying query length and based on factoid and non-factoid queries. We show that some of the methods significantly improve the rankings in the education domain.

1. Introduction

Students choose the Internet as their primary resource for research when it comes to finishing a homework or learning a new subject in their curriculum whether the subject at hand is a Mathematics problem or an open-ended Social Sciences project. Web search engines enable students to scan a variety of web pages and provide a list of possible candidates for them to choose from. It is up to the student then, to extract the information available in the returned web pages. However, most of the available web search engines are designed for general purpose and may not serve the students' needs that are very specific to the education context. For example, the web pages returned to the queries can be from the other topics due to a diversification process employed by the search engine or just because of textual similarity. This leads to the formation of domain-specific or vertical search engines. The advantages of vertical search engines over general purpose ones include the specificity of results in the concerning domain (e.g., education) due

* Corresponding author.
E-mail addresses: tolga.yilmaz@bilkent.edu.tr (T. Yilmaz), rifatozcan1981@gmail.com (R. Ozcan), altingovde@ceng.metu.edu.tr (I.S. Altingovde), oulusoy@cs.bilkent.edu.tr (Ö. Ulusoy).

https://doi.org/10.1016/j.ipm.2018.10.013
Received 24 February 2018; Received in revised form 7 October 2018; Accepted 16 October 2018
Available online 24 October 2018
0306-4573/

to a narrower scope (Chau & Chen, 2003) and it is easier to customize user experience in the sense that the results can be sorted more easily with respect to their context (Diligenti, Gori, & Maggini, 2002). Additionally, if we approach search as a learning process, general web search engines serve the need for factual information learning and there is still room for improvement in other learning types such as understanding, analysis, and applications of the learned information (Rieh, Collins-Thompson, Hansen, & Lee, 2016). Additionally, analysis by Usta, Altingovde, Vidinli, Ozcan, and Ulusoy (2014) reveals that educational search behavior differs from general web search in terms of query characteristics (e.g., the high percentage of repetitive queries by students) and click behavior (e.g., more clicks on pages containing active content such as animations).

The research by Purcell et al. (2012) suggests that teachers say teens are 94% “very likely” to use general search engines but only 26% of them are rated as excellent or very good on formulating effective search terms. The ratio of full sentence questions or natural language usage increases as the age of children drops. Bilal (2002) reports that 35% of 7th-grade children uses natural language queries and Schacter, Chung, and Dorr (1998) inform that 65% of 5th to 6th-grade children use such queries. Tu, Shih, and Tsai (2008) report a similar behavior for 8th graders. Kammerer and Bohnacker (2012) argue that not only children tend to use natural language more frequently while searching information but also the use of natural language is beneficial for them. In their work, children were assigned tasks to solve doing Internet searches. The ones who use question-like natural language queries were able to solve the tasks easier.

Student behavior, besides web search, includes asking for help in online communities (Purcell et al., 2012). These websites in the last decade evolved from forums with continued discussion to question and answer (Q&A) websites which are more focused on the quality of questions and answers. Although forums are still popular as discussion platforms, it is easier to extract information from structured community Q&A websites. Additionally, while search engines list relevant web pages, one needs to scan these pages and extract the answer from them. In Q&A websites, there is direct access to answers. General purpose Q&A websites dominate the World Wide Web today such as Quora (2018), Yahoo! Answers (Compete Site Analytics, 2015) and StackExchange (Stack Exchange, 2018). There are many small-scale homework sites and also subsections under popular websites such as Yahoo! Answers Education & Reference Category (Yahoo! Answers, 2018).

Turkish student community also populates educational Q&A websites. Brainly.co (Brainly Inc., 2018) operates an educational Q&A platform in 35 countries including Poland, Russia, Brazil, the U.S.A., Spain, France, and Turkey. They report 60 million monthly unique users across their websites. Turkish one is named as EOdev.com (EOdev, 2018) and it claims to have over 4.5 million questions.

Students at primary and secondary school age may not form effective search terms and generally, submit their queries in the form of natural language queries. General purpose search engines try to produce diverse results for such queries based on query keywords without considering the educational intent or subject of the query. We showed that 40% of irrelevant results for such educational queries in a large-scale commercial search engine is due to the mismatch between the educational subject (e.g., Math, Science) of queries and results (see Section 5.1). Therefore, it is important for a search engine to understand the educational subject of queries so that it can prevent such irrelevant results. The following summarizes our main contributions to tackle the problem.

- We build an educational subject classifier for queries. There is no publicly available dataset for such queries in Turkish. Therefore we utilize educational questions posted in educational Q&A websites and manually label them with educational subjects. Since students also tend to submit natural language queries, questions can be safely regarded as queries.
- We utilize a diverse set of lexical, syntactic and semantic features, query results from a search engine and query expansion for classifying educational subject of queries at a resulting 83.58% accuracy.
- We propose point-wise and list-wise re-ranking mechanisms that optimize the ranking order based on predictions of the educational subject classifier. Two of our re-ranking methods achieve statistically significant results in Normalized Discounted Cumulative Gain (NDCG) metric compared to the original ranking obtained from a large-scale commercial web search engine.
- To the best of our knowledge, this is the first manuscript that works on Turkish question answering and search engines in the educational domain with this number of features and datasets. Additionally, we implemented some features for Turkish that are already available in English and there is no obvious obstacle to the implementation of our general solution for other languages. For some features that have good implementations in English but not in Turkish (e.g., headwords) we used replacement features (e.g., object and subject) and derived other features from them (e.g., object and subject phrases, semantic features).

The paper is structured as follows. In Section 2, we give a brief review of the literature related to our work. Section 3 presents our educational question classification method and Section 4 gives the experimental classification results. Section 5 motivates the use case of our classifier in ranking results of educational queries. Section 6 explains the proposed ranking methods. We evaluate these methods in Section 7. Finally, Section 8 concludes our paper.

2. Related work

In this section, we briefly discuss the literature related to Question Answering (Q&A) systems, query and question classification, and educational content ranking.

2.1. Social question answering

Social Q&A websites have gained worldwide popularity as a new way to access information for people who seek answers to their

questions. Yahoo! Answers website was visited by 46 million single users in May 2015 (Compete Site Analytics, 2015) and StackOverflow.com (Stack Overflow, 2018), an online Q&A community for programming questions, has over 4 million users.

According to Gazan (2011), research on Social Q&A can be divided into two categories as question quality/classification and answer quality/classification. Example works include research on finding better users in the community (Pal & Konstan, 2010; Yang, Tao, Bozzon, & Houben, 2014), identifying good questions (Li, Jin, Lyu, King, & Mak, 2012; Liu, Shen, & Yu, 2017), answers (Harper, Raban, Rafaeli, & Konstan, 2008), finding the intent behind the questions (Chen, 2014; Harper, Weinberg, Logie, & Konstan, 2010), and classifying questions based on type (Harper, Moy, & Konstan, 2009) (i.e., is it opinion asking or fact seeking), based on subject (Cao, Cong, Cui, Jensen, & Zhang, 2009), based on detailed subtypes (Wu, Hori, Kashioka, & Kawai, 2015) and based on motivation (Espina & Figueroa, 2017).

In the education field, Ghosh and Kleinberg (2013) focus on K-12 forum sites as opposed to Q&A sites and try to find what attracts users to contribute. They also point out the differences between forums and Q&A sites under the assumption that forums are assisted by actual instructors. Mao (2014) focuses on the attitudes of high school students for social media. Their data show that for social media to be an effective learning environment, the extraction of current social media habits of students, the adoption of these into such learning environments, and interacting with students are necessary. Gurevych, Bernhard, Ignatova, and Toprak (2009) describe how to use social media for educational question answering. They use a classifier for the subjectivity of questions. They also present a Q&A system that fetches answers from social media content.

2.2. Question classification

Question classification is an important part of question answering systems and lately Community Question and Answer (CQ&A) websites, and there is a considerable amount of work in the last decade. Question classification generally serves as an intermediate step towards achieving better results in question answering systems.

In one of the earlier works, Zhang and Lee (2003) use machine learning techniques to create more robust classification systems. They try multiple learning algorithms that take n-grams as features and report SVM to be the most accurate. Li and Roth (2002) use semantic and syntactic features with SnoW algorithm to classify questions. Metzler and Croft (2005) similarly work on semantic and syntactic features and on different data sets composed of fact-based questions. They show that question classification using machine learning, SVM in their case, is robust compared to rule-based classifiers which require an immense amount of effort.

Huang, Thint, and Qin (2008) introduce headwords and their hypernyms as an important feature of question classification. Most of the later works include this feature as well (Loni, 2011a; Mishra, Mishra, & Sharma, 2013; Silva, Coheur, Mendes, & Wichert, 2011). Most recent features and a broad survey are provided by Loni (2011b).

There are works in classifying educational questions as well. Li, Samei, Olney, Graesser, and Shaffer (2014) classify questions in an epistemic game that tries to teach various concepts to students with instructor involvement. Sangodiah, Muniandy, and Heng (2015) give a short review of the existing work on educational question classification featuring Bloom's taxonomy (Bloom & Engelhart, 1969) which is sorting educational objectives into hierarchical levels of complexity and mastery. Research on this field includes the works that employ algorithms such as SVM (Yahya & Osman, 2011), rule-based classifiers (Haris & Omar, 2012) and neural networks (Yusof & Hui, 2010).

The work in Figueroa and Neumann (2016) bases itself on question-like search queries from Yahoo! Answers. They try to motivate the connection between search engines and CQ&A like we do. They utilize many features and external data sources. In Table 1, we show some of the works mentioned above with their features and algorithms, and the method we use in this work. Our work, compared to most of the other work, utilizes external resources such as books, term collections, and search engine result pages. We also use an ensemble classifier rather than using a single algorithm. We base our work on the education context whereas most question classification research is based on general categories.

2.3. Educational content ranking

Some of the recent works on search engine result page ranking employ Learning to Rank (LETOR) algorithms. For example, in the education domain, Usta et al. (2014) analyze the search log of an educational search engine and Usta (2015) makes use of a LETOR technique that utilizes textual similarity, query, and document specific, session-based and click based features to improve the ranking of a commercial educational search engine. Similar works mostly focus on ranking learning objects. In this context, Ochoa and Duval (2008) define relevance metrics that take into account multiple levels such as topical, personal and situational relevance. Recommending learning objects is another topic that attracts interest. Verbert et al. (2012) provide a survey on context-aware recommender systems. Our work attempts to improve ranking by using classification results. In our work, we rank web pages that are not necessarily educational, which is different from the other works involving ranking learning objects that are known to be educational. Eustace, Wang, and Li (2014) also tackle the noisy result pages by applying Singular Value Decomposition on the result set. Different from their work, we also use an explicitly trained classifier to re-rank result pages.

3. Educational question classification

We approach the question classification task first as a text classification problem. We follow the bag-of-words model so that terms are considered as features. Later, we introduce other features special to question classification which can be seen in the next subsection.

Table 1
The list of works on question classification by features and algorithms.

Author	Features	Algorithms	Language
Hermjakob (2001)	N/A	Parse Tree	English
Li and Roth (2002)	Words, POS Tags, Chunks, Named Entities, Head Chunks, Semantically Related Words	SnoW	English
Zhang and Lee (2003)	N-grams	Naive Bayes, Winnow, Decision Tree, SVM	English
Metzler and Croft (2005)	Bag-of-words, N-grams, POS Tags, Question Word, Noun Phrases, Question Length, Long Distance k-grams, Headword, Hypernyms of the head word using (WordNet)	SVM	English
Li and Roth (2006)	Bag-of-words, POS Tags, Head, Named Entities, WordNet, Class-specific Related Words, Distributional Similarity of Words	Winnow	English
Huang et al. (2008)	Wh Words, Headwords, WordNet, Direct-Indirect Hypernyms, Unigrams, Wordshapes	SVM, Maximum Entropy	English
Silva et al. (2011)	Unigrams, Headwords	Rule Based, SVM	English
Loni (2011a)	Unigrams, Bigrams, Headwords, Headrules, Word Shapes, Wh Words, Hypernyms, Query Expansion, Question Category	NN, SVM, LSA	English
Mishra et al. (2013)	Unigrams, Wordshapes, Headword, Question Patterns, POS Tags, Hypernyms, Question Category	KNN, Naive Bayes, SVM	English
Li et al. (2014)	POS Tags, Wh Words, Categorical Wordlists	Decision Tree	English
Vlasák (2015)	Bag-of-words	KNN, Naive Bayes, SVM	Czech
Figueroa and Neumann (2016)	Bag-of-words, Latent-topic models, Acronyms, String Analysis, String Distances, POS Tags, Yago2s, Wikipedia, Yahoo Categories	Maximum Entropy, Winnow	English
Our work	Unigrams, Bigrams, Question Length, Wh Words, Word Shapes, POS Tags, Tagged Unigram, Word Dependencies, Object and Subject, Object and Subject Phrases, Hypernyms, Hyponyms and Other Semantic Relations, Query Expansion, Search Engine Results	Ensemble of SVM (Linear RB), Complement Naive Bayes, Naive Bayes, Discriminative Naive Bayes, Hyper Pipes and Search Based Classifiers	Turkish

Fig. 1 shows the overall design of our question classifier. Questions are first preprocessed and then a query expansion mechanism is applied to further enhance the question content. Then various features (the last row in Table 1) are extracted and several supervised classifiers (e.g., Naive Bayes, SVM) are learned. We also design search based classifiers relying on course textbooks, education term collections. Then all these classifiers are ensembled. In the next stage, we incorporate search engine result pages into our ensemble classifier to further improve our accuracy. The method is basically as follows:

1. Pre-classify questions and their query results with the ensemble question classifier.
2. Reclassify questions with a weighted majority voting scheme based on pre-classified questions and their query results.

Even though our ultimate aim is to improve the ranking of results for educational questions, we can still rely on initial results obtained from search engines since most of the time, the majority of the results are consistent with the subject of the educational query. Therefore, we believe that we can improve our classifier accuracy using this extra information and re-rank the results so that results referring to non-educational context can be eliminated by the predicted question subject.

In the following subsections, we first describe our data sets (Section 3.1) and (then present the details of our features (Section 3.2). In Section 3.3, we describe our ensemble method to combine various classifiers.

3.1. Data sets

We utilize the content of a Turkish educational Q&A website, textbooks that were taken from Turkish Ministry of Education website[1], educational objects from Vitamin education service (Vitamin Eğitim, 2018), and query results obtained using Bing API (Microsoft Azure, 2018) in response to educational questions. Below, every data set is explained in detail.

3.1.1. Educational social q&a website questions

We have collected the entire Q&A data from msxlabs (msxlabs, 2018). This website has an interface very similar to StackOverflow; there are questions, comments, and answers. The reasons behind choosing this website as a resource include the continued moderation of the content, its content being in coordination with the current curriculum of the education system, its active user base and its structural similarity to well-known Q&A websites. The dataset contains 10,717 unlabeled questions. Although question bodies are useful in increasing the accuracy of classification task such as finding out informational / non-informational questions as in (Palomera & Figueroa, 2017), in our work we do not use them since they cannot be issued as search engine queries. Additionally, the

[1] https://www.meb.gov.tr/.

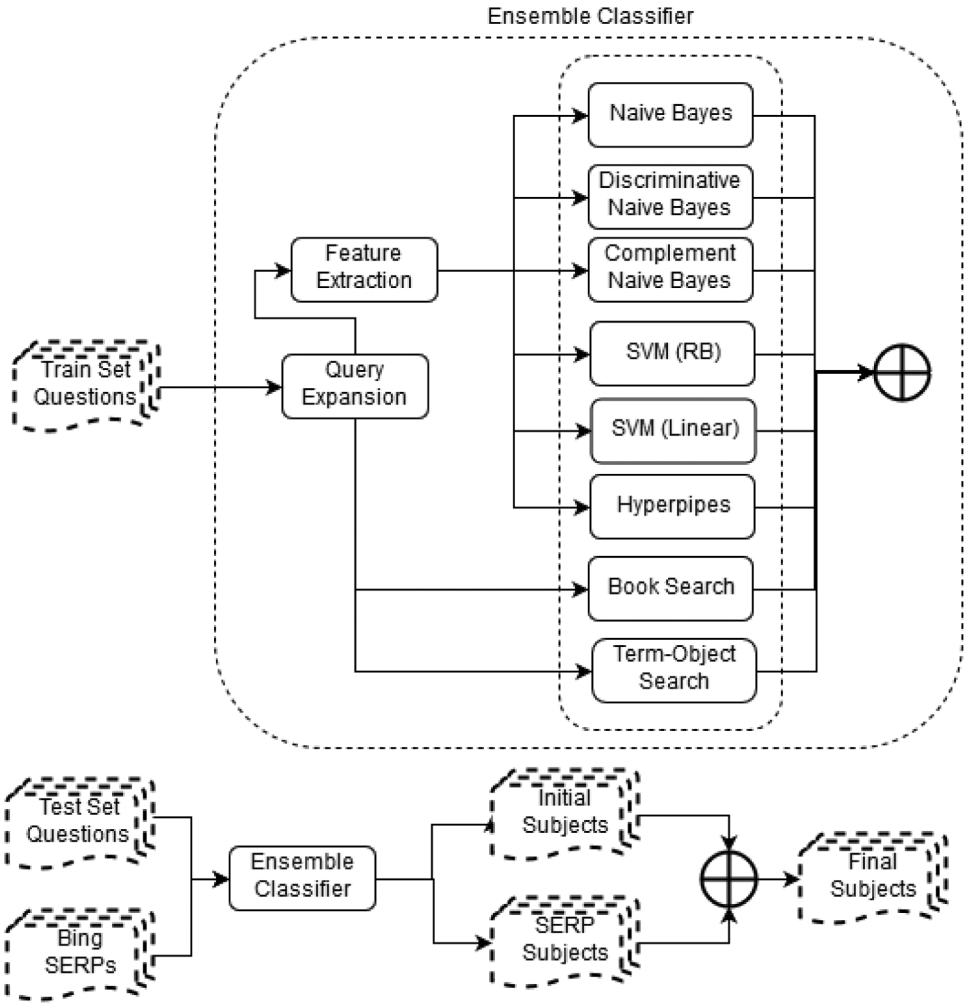

Fig. 1. Overview of the Educational Question Classifier.

questions in our dataset are mostly comprised of just titles, few questions had bodies. Almost all of the questions were in the natural language question form. We did not do extra processing to correct or eliminate questions with spelling errors. The number of questions with spelling errors was small due to moderation.

3.1.2. Textbooks

Textbooks for students are great resources to find answers to questions. We have downloaded the pdf versions of the textbooks from the Turkish Ministry of Education website for every grade in Turkish Education System for the subjects: Turkish, English, Mathematics, Republic History, Science, and Social Sciences. Then, we have extracted the content of these pdf textbooks and merged the books from different grades with the same subject, leaving each document representing a corpus for a subject. In the end, there are 7 large documents.

3.1.3. Online course content

Vitamin web service (Vitamin Eğitim, 2018) is a paid educational tool in Turkey. In their system, there are learning objects representing tiny bits of information in different subjects. A learning object (see Table 2) consists of a title (i.e., an educational topic), a description and a path (i.e., subject such as Math and Science, grade information such as 8th grade). For us, they serve as mappings

Table 2
Example learning object .

Title	Description	Path
Reflection and Refraction of Light	Reflection and Refraction of light is described using a daily example.	Science / 7th Grade / Light / Reflection and Refraction of Light /

from topics to subjects. We think of topics as queries. We have 6264 unique objects like this. Although not extensive, this content was prepared by the experts of the company and is, therefore, reliable after some preprocessing.

3.1.4. Educational term collection

We crawled a list of term-definition pairs from (Dersimiz.com, 2018) where terms are educational entities such as "mitochondria: an organelle... ". Our compiled list contains about 33K terms. For each subject, the numbers of terms are: Mathematics (671), Science (7071), Miscellaneous (14,870), Social Sciences (6974), Turkish (1552), Religion (1430), Republic History (1170).

3.1.5. Bing query results

We have used the Bing API to collect top 50 query results for a randomly selected set of 5000 questions from the msxlabs data. Considering that some queries returned less than 50 results, we have approximately 240,000 query results.

3.2. Query expansion

We implement a query expansion (QE) mechanism to enhance the query classification. The idea is to expand the question with terms from the same subject in order to increase the accuracy of the training and testing. Rather than using traditional query expansion mechanisms such as Rocchio's or finding the headword of the query then expanding it, we use a simpler approach taking advantage of the structure of our term list. Using the term list we compiled, we create an index in the form <title, definition, subject> using Lucene.[2] Taking the question as a query, we search through our index finding the most similar documents. Since the title itself intuitively represents the most important word in that document, we simply take the title as our expansion term. However, further improvements are necessary. We calculate the score of each document first and use an empirical term frequency-inverse document frequency (TF-IDF) threshold of 0.8. Documents passing the threshold are taken into account. Furthermore, before expanding a query we ensure that the expansion term is not a rare term. The classification error induced by falsely expanded queries with rare expansion terms may be larger than the ones with more common expansion terms. For this reason, we simply check whether the terms document frequency is larger than a threshold (3 in this case). We also find out that a larger number of expansion terms degrades the classification and use only one expansion term. Due to the limited size of the term list, more specific educational terms could not be retrieved. If there was a larger educational term dictionary, more expansion terms could have been retrieved and this could have been useful. Finally, since we know the subjects of the expansion terms as they are indexed, we do a premature classification of the input query using a voting among the returned documents and find an initial subject. Documents that do not belong to this subject are discarded. After all these improvements, we observe that only around 14% of queries are expanded. The empirical thresholds are determined in the training phase of the classification. Below, we give some examples of questions and resulting expansion terms after applying our expansion method.[3]

- Bilgisayar ağları nelerdir?—Internet
  What are computer networks? — Internet
- Dünyanın şekli tam olarak nasıldır?—Geoit
  What is the exact shape of the Earth? – Geoid
- İğ ipliği oluşturma görevi hangi organele aittir—Sentriyol
  Which organelle is responsible for creating interconnecting fibers?–Sentriole

After finding the expansion term, we append the term at the end of the question in order to make it more specific. In our query expansion experiments, we observed that we were able to provide hints, if not exact answers, to the posed questions.

3.3. Features

The features of question classification are divided into three categories: lexical, syntactic and semantic. We mostly benefit from the survey in Loni (2011b). The features we implement are presented in Table 3. As we show in Table 1, compared to the other studies in the literature, our work covers an extensive list of lexical, syntactic and semantic features for this task and apply various classification algorithms. Turkish is a morphologically complex language (Eryiğit, Nivre, & Oflazer, 2008). There are many efforts for various Natural Language Processing tasks for Turkish but they are not as mature as they are in English. A great article on the challenges of Turkish NLP is (Oflazer, 2014). We would access English resources more easily if our work and datasets were in English only. For this reason, we had to remove some of the features because we were either unable to retrieve or integrate into our study. One example of that is the Turkish WordNet (Bilgin, Çetinoğlu, & Oflazer, 2004). Other examples would be keyword extraction and named entity recognition in Turkish for which we had hopes to use their results as features. Additionally, we had to implement the retrieval of some of the features such as object, subject, object phrase and subject phrase. In this section, we give the details of each of the utilized features.

[2] http://lucene.apache.org/core/.

[3] The example questions are provided in their original form (i.e., in Turkish), followed by their English translation.

Table 3
Features and abbreviations .

Type	Feature
Lexical	Unigrams
	Bigrams
	Trigrams
	Wh words
	Word shapes
	Question length
Syntactic	POSTags
	Fine Grained POSTags
	Tagged Unigrams
	Object and Subject
	Object and Subject Phrases
	Word Dependencies
Semantic	Synonyms
	Antonyms
	Hypernyms
	Hyponyms
	Side Concepts
	Associated Words
	Similar to
	Used for
	By goal

Table 4
Length statistics for each subject.

Subject	Median	Mean	Std. Dev.	Skewness
Turkish	7	13.07	44.81	17.49
Math.	10	14	15.17	8.31
English	6	13.64	20.49	4.43
Rep. Hist.	8	13.78	36.37	9.21
Religion	8	10.61	11.95	6.76
Soc. Sci.	7	9.82	18.31	15.13
Science	7	11.30	41.62	16.18
Misc.	6	11.71	40.97	16.48

3.3.1. Lexical features

Lexical features are simply the words in a question rather than the grammatical structure of it.

Unigrams, Bigrams, and Trigrams Tokenization of a sentence or a word results in the n-grams of the sentence or the word. Thus, n-grams are either words or characters comprising the sentence or the word. In this work, we use word n-grams; namely, unigrams, bigrams, and trigrams.

Wh-words Wh-words or question words describe what is wanted about the subject of the question and therefore may be useful in determining the category of the question. In English, *Who, Where, What, Which, When, How, Why* are the most commonly used wh-words. For Turkish, we determined the following words as question words: Ne (*What*), Nere (*Where*), Nasıl (*How*), Neden (*Why*), Kim (*Who*), Hangi (*Which*), Ne Zaman (*When*), Kaç (*How many*). Since a question can include one or more of these words, we represent this feature as <featurename, occurence>. Consider the question "2020 Yaz Olimpiyatları nerede ve ne zaman yapılacak?" (*When and where will the 2020 Olympics take place?*). Its extraction is performed as follows:

{(Ne, 0), (Nere, 1), (Nasıl, 0), (Neden, 0), (Kim, 0), (Hangi,0), (Ne Zaman 1), (Kaç, 0)}

{*(What, 0), (Where, 1), (How, 0), (Why, 0), (Who, 0), (Which, 0), (When, 1), (How many, 0)}*

Word Shapes A question can include words in many forms such as *lowercase, uppercase, all digits, mixed* and *other*. For the same question, these features are extracted as:

{(All digit, 1), (Lowercase, 5), (Uppercase, 2), (Mixed, 0), (Other,0)}

Question Length The length of the question in terms of words may be an indicator of its category as it is used in some research such as (Metzler & Croft, 2005). When we investigate the question lengths of each category, we see that there are some subtle differences. It might be the case that longer questions are from a certain category. For instance, Mathematics questions have a larger median number of words than any other category. It has the second smallest standard deviation (see Table 4).

3.3.2. Syntactic features

Syntactic features represent the grammatical structure of a question. We start building this type of features using the ITU NLP API[4]

[4] http://tools.nlp.itu.edu.tr/

Table 5
Dependency information of "What are the distinctive properties of pure substances?" question (*"Saf maddelerin ayırt edici özellikleri nelerdir?"*).

Id	Form	Lemma	Cpostag	Postag	Feats	Head	Deprel
1	Saf	saf	Adj	Adj	_	2	MODIFIER
2	maddelerin	madde	Noun	Noun	A3pl|Pnon|Gen	8	SUBJECT
3	ayırt	ayırt	Noun	Noun	A3sg|Pnon|Nom	4	MWE
4	_	et	Verb	Verb	Pos	5	DERIV
5	edici	_	Adj	Agt	_	6	MODIFIER
6	özellikleri	özellik	Noun	Noun	A3pl|Pnon|Acc	8	OBJECT
7	_	ne	Pron	Ques	A3pl|Pnon|Nom	8	DERIV
8	nelerdir	_	Verb	Zero	Pres|A3sg|Cop	0	PREDICATE
9	?	?	Punc	Punc	_	8	PUNCTUATION

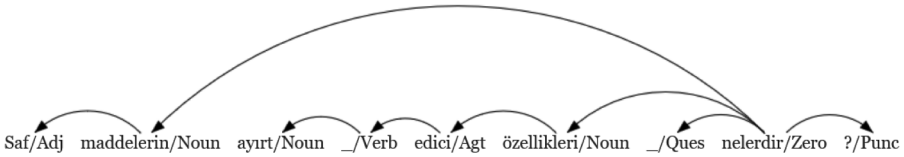

Fig. 2. An Example Dependency Tree.

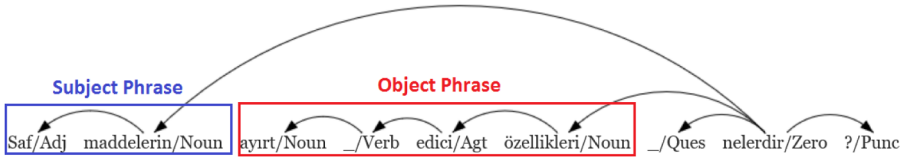

Fig. 3. Subject and object phrases in a dependency tree.

described in (Eryigit, 2014). The extraction of syntactic features depends on the construction of the dependency tree of a given question. Consider the question: "Saf maddelerin ayırt edici özellikleri nelerdir?" (*What are the distinctive properties of pure substances?*). ITU API generates the syntactic information in Table 5 using CONLL (Buchholz & Marsi, 2006) format. Using the binary dependency information, one can generate a dependency tree as in Fig. 2.

Part of Speech Tags We used part of speech (POS) tags and represented this feature as bag-of-postags. An example based on the question in Table 5 is given below. We also consider the coarse and fine grained POS tags.

{"Adj Noun Noun Verb Adj Noun Pron Verb Punc"}

Tagged Unigrams This feature combines the unigrams with their corresponding pos tags. An example based on the question in Table 5 is given below.

{"saf_Adj madde_Noun ayırt_Noun et_Verb -_Adj özellik_Noun ne_Pron -_Verb ?_Punc"}

Object and Subject As shown in Table 1, headwords and its derivations are commonly used in question classification. However, we were unable to find an implementation of a direct headword extraction algorithm for Turkish. Nevertheless, we needed a replacement for headwords to generate some other syntactic features and all semantic features. As a result, we chose Object and Subject to be the easiest replacement for headwords.

The object and subject can be extracted using the dependency structure of a sentence. Analyzing languages with dependency grammar is useful with free word order languages such as Turkish. Dependency parsing has also been studied for Turkish (Eryiğit et al., 2008). It is also available using the ITU API. Using the CONLL format we easily extract the object and subject. For 29% and 73% of the questions, we are able to extract the object and the subject, respectively. Only 15% has both the object and the subject. The following is an example of the object and subject information based on the question in Table 5:

{"madde özellik"} *{"substance property"}*

Object and subject phrases After finding the object and subject of a question, we want to expand them to extract more information. The object and the subject are single words but they may be a part of a phrase. If we extract the subject of the question "What do the green plants do to generate energy" as "plant", we miss the phrase "green plant", which may be more informative since it is exactly what the question is about. To find the object and subject phrases, we first find the object and subject and get their dependents recursively. Considering the dependency information in Table 5, "madde" (*matter*) is the *SUBJECT*. It has one dependent, which is "saf" (*pure*). Similarly, "özellik" (*property*) is named as *OBJECT*. It also has a dependent at the 5th token whose dependent is the 4th whose dependent is the 3rd. By combining these, the object and subject phrases are found as "saf madde" (*pure substance*) and "ayırt et özellik" (*distinct property*), respectively. These are much better in capturing the essence of the question. We also show the corresponding information in Fig. 3.

{"saf madde ayırt et özellik"} *{"pure substance distinct property"}*

Word dependencies By using the dependency tree we constructed, we extract one to one dependencies and use them as bag-of-

Table 6
Semantic features.

Feature	Description	Example	Ext. %
Synonyms	Other ways of saying a word	*autumn* → *fall*	65
Antonyms	Semantic opposites of a word	*autumn* → *spring*	1.6
Hypernyms	Generalizations of a word	*autumn* → *season*	27
Hyponyms	Specifications of a word	*autumn* → *october*	1.6
Side Concept	Siblings of a word	*autumn* → *summer*	1.1
Associated Words	Words that do not have a family relationship but somehow related	*observation* → *observation satellite*	6.6
Similar to	Lexically similar	*observation* → *observer*	1
Used For	What the source concept is used for	*eraser* → *erasing*	0.7
By Goal	This relation is quite similar to the *Used for* relation but the target concept is not a direct consequence of the source concept	*assembly* → *talking*	0.7

dependencies, i.e., they are depth-1 parent-child relations.

{"saf-madde madde-ne ayırt-et et-özellik özellik-ne ?-ne"}

{*"pure-substance substance-what be-distinct be-property property-what ?-what"}*

3.3.3. Semantic features

We extract semantic features based on the objects and subjects of questions. In order to extract semantic features, generally, WordNets are used. We use the semantic relation dataset from (ITU Kemik, 2018). It contains 127,203 relations and is general-purpose not educational. The data set was constructed using 98,107 words from TDK (Turkish Language Association) dictionary (Türk Dil Kurumu, 2018) and 160,049 from Vikisozluk (Vikisözlük, 2018) which is the Turkish version of Wiktionary (Wiktionary, 2018) which is a collaborative multilingual dictionary project. It has "word ← relation → word" format. Based on the data set, the following features are extracted and used as bag-of-words. We also list the ratio of questions that we were able to extract for each feature. Here, we first extract the question subject and object if possible (as stated in the previous section) and then use them as keys to look up the features in Table 6.

3.4. Ensemble method

We implemented an ensemble method that utilizes simple majority voting over multiple classifiers. It should be noted that by majority we actually mean first-past-the-post rule, i.e., the subject with the highest number of votes directly wins. Eq. (1) gives the rule where q represents the question, c_i the classifier i from set $C = \{c_1, c_2, ..., c_n\}$ where n is the number of classifiers. Eq. (2) shows how the probability of assigning q to each subject is calculated where subject s is from the set $S = \{s_1, s_2, ..., s_8\}$ (i.e., subjects from Table 4).

$$Classify(q) = mode\{c_1(q), c_2(q), ..., c_n(q)\} \tag{1}$$

$$P(s|q) = \frac{\sum_{c \in C} P(s|c, q)}{\sum_{\hat{s} \in S} \sum_{c \in C} P(\hat{s}|c, q)} \tag{2}$$

We have also experimented with other ensemble techniques such as Adaboost, Random Forests and Bagging or using a neural network. However, the results were not satisfactory. We could also implement an ensemble of ensembles but that would make already slow ensembles slower, making it harder to experiment. Our simple implementation outperformed such more advanced techniques. This may be due to the size and the distribution of the data set. In our case ensemble techniques that use similar classifiers over imbalanced data did not outperform our simple technique.

3.4.1. Preprocessing

We applied traditional information retrieval techniques in the preprocessing step of the training and test sets. A stemming tool[5] is applied to reduce the number of features. We also remove the Turkish stop words (Can et al., 2008) and do lowercasing and noise removal. Here, noise is defined as character sets (for a search query) such as emoticons and punctuation and is removed with simple filtering.

3.4.2. Search based classifiers

These classifiers utilize the resources we have and do not require training. They are also very fast. For our search based classifiers, we utilize Lucene.

Book search classifier We have implemented a searcher on the book set using Lucene. First, we indexed the books. If we search a question in these books/documents, since each book/document represents a subject such as Math and Science, the first matching

[5] https://github.com/hrzafer/resha-turkish-stemmer.

document for this question becomes the classified subject. Here, Lucene's default similarity is used and a threshold t = 0.025 is applied to further improve the accuracy. Documents passing the threshold are taken into account. The value of the threshold is determined empirically in the training phase and it is highly dependent on the similarity measure.

Object search classifier We retrieve the subject information such as "Science" from the path column of a learning object given in the format (title, description, path). We treat each object as if it is a document and we want to search through these objects. Thus, we index them (i.e., their title + description) along with their subjects. If a question is asked, we search through our learning object index and return the most similar objects (i.e., documents) in our case. Since the index returns the instances based on a ranking, doing a weighted/graded voting may be better than the equal-weight majority voting. We settle with the following voting function based on our observations. Here, r represents an object returned as a result and s is a subject.

$$\gamma(r, i) = \begin{cases} 1 & : subject(r) = s \\ 0 & : subject(r) \neq s \end{cases} \tag{3}$$

$$Classify(Q) = arg\ max_s \sum_{j=1}^{k} (\gamma(r_j, s)/(j+1)) \tag{4}$$

For the value of k, the number of results to retrieve, we decided on empirical value 40. If we had simple majority voting, k = 5 would be our choice. But we decided on the weighted voting because it performs better based on our observations. Furthermore, we use BM25 here without any threshold.

Term search classifier We indexed the terms in the educational glossary where a document is represented as <term, definition, subject>. Using the same idea in the object search, we implement a classifier. The weighting scheme is the same as the object search, however, we only take the documents passing the 0.9 BM25 score into account. This value is determined during training.

Term-object search classifier We combined the data from two data sources of educational glossary and Vitamin learning objects and build a search classifier that searches in this combined index. This classifier employs the same threshold and weighing scheme as the Term Search classifier.

3.4.3. Machine learning classifiers

In our experiments, we have used Weka[6] with the following classifiers: Naive Bayes Multinomial Updateable (NB), Complement Naive Bayes (CNB), Discriminative Naive Bayes (DMNB), SVM (Linear, Radial Basis(RB)), and HyperPipes (HP). For all these classifiers, we have tried to optimize their specific parameters including normalization and the number of words to store.

3.5. Exploiting search engine results

Search Engine Result Pages (SERPs) are the web pages displayed by a search engine as a response to a query. In our case, it represents the title, link, and description of each page. Since we expect the majority of SERPs to be relevant to the question most of the time, it is intuitive to use SERPs for classification. In fact, if we had a query result classifier, we may have expected that the majority of results returned to the query to be from the same subject. Since building a new classifier for SERPs requires extra effort of labeling, we use our question classifier to classify the SERPs as well. Since SERPs are in a different form than questions, we only use the bag-of-words features for classifying them and do not use the syntactic and semantic ones.

For finding out what the majority of query results returned to a query says about the subject of the query, we use a voting scheme that emphasizes the ranks and per-subject probabilities of query results and initial classification of the question. In this way, top documents will have a higher impact on the voting (see Eq. (5)). Here, s_q represents the final subject of the query. d_j is a query result returned to query q at rank j. $P(s|q)$ and $P(s|d_j)$ are the initial probability of query q and query result d_j being in subject s, respectively. These are calculated using our ensemble method (see Eq. (2)). We also optimize the classification accuracy by selecting the top 20 results.

$$s_q = arg\ max_s \left[\lambda P(s|q) + (1-\lambda) \sum_{d_j \in D} \left[\frac{P(s|d_j)}{\log_2(j+1)} \right] \right] \tag{5}$$

λ is a confidence parameter between 0 and 1. When it is bigger than 0.5, we put more trust in the initial classification of the query. When it is smaller than 0.5, we trust more on the knowledge of query results. It is an empirical value and can change according to the data set. If we are to use another search engine or trust our classifier more, the value of lambda would possibly change.

In order to incorporate trust, we use the assumption that questions with fewer terms are less informative about their subjects. Therefore, in classifying shorter questions, we trust the voting among query results more with smaller λ. This results in the following setup: For shorter questions (i.e., shorter than or equal to 6 tokens) $\lambda = 0.3$, otherwise $\lambda = 0.5$. The choice of "6" is not arbitrary. It is the median number of tokens of a question in our data set. Additionally, using this method we are able to calculate the final subject probabilities $P(s_i|q)$ of queries without needing a transformation, which can be used for various purposes later.

We use the snippets of query results when classifying them. Other approaches may include aggregating all query results and

[6] https://www.cs.waikato.ac.nz/ml/weka/.

Table 7
Subject distributions as result of labeling.

Social sciences	1463	English	148
Miscellaneous	795	Mathematics	526
Religion	231	Science	1125
Turkish	511	Republic history	201

making a classification on this aggregated document. Utilization of full web pages was also possible. We also tried those and chose the previously stated individual weighted voting scheme over these other approaches because those methods did not show promising results.

4. Classification results

We have randomly selected 5000 questions from our question set and labeled them based on the classes in Table 7 which also shows the subject distribution. Random sampling is used because we do not know the class distribution in the original data set (described in Section 3.1.1). The labels were given by three annotators (the first author and two other persons outside the work). The first annotation was done by a single annotator alone. The second annotation was done by the other annotators each labeling separate halves. When there was a tie, the annotators would reconsider the category of the question together. The kappa score for measuring the inter-annotator agreement for the initial annotations was 0.81. The decision on this number of subjects is in accordance with the external resources; i.e., not all datasets shared the same subject labels and our setup is a combination that made the use of the external data sets possible.

In our experiments, we obtain the accuracy by 10-fold cross-validation. For each fold, we train 90% of questions and classify 10% of questions and query results returned to them. Then, we improve the classification of these pre-classied questions using the pre-classified query results by employing the voting technique described in the previous section.

As the feature space, we first try the bag-of-words model, namely unigrams and bigrams. According to this setup, Table 8 gives the results of individual classifiers, our ensemble method and the SERP enhancement over it. The ensemble method contains NB, HP, CNB, DMNB, Book Search, Term-Object Search and SVM (Linear and RB). The results show that Naive Bayes and SVM (Linear) based classifiers perform well but our retrieval based classifiers perform the worst. We also see the small but positive effect of query expansion which is 0.2% over the ensemble. SERP enhancement is observed to be 4.8% over this. Query expansion and SERP enhancements add up to 5% improvement.

Next, we run experiments on a broader feature space and try to find the features that are the most beneficial. Table 9 shows our greedy approach reducing the feature space. Our approach is similar to the sequential backward selection (SBS) algorithm where we start with all the features and eliminate those that provide the highest accuracy gain when removed at each step. Jain and Zongker discuss the advantages and limitations of this and other feature selection algorithms such as sequential forward selection (SFS) in Jain and Zongker (1997). As the results suggest (the last row in Table 9), in addition to unigrams and bigrams, we find tagged unigrams, object, subject, hypernyms of the object and subject to be the most beneficial. Some of the lexical features such as question length, wh-words, and word shapes were not beneficial and thus not included in the final model although there were studies showing the opposite. This may be due to the language or data set differences.

Table 10 gives the confusion matrix for the final classifier which is built using unigrams, bigrams, tagged unigrams, object, subject, hypernyms of object and subject, query expansion and SERP enhancement.

During the labeling process, we put questions that do not fit into any other category but still are education-related into the Miscellaneous category. Thus this category contains the most unique questions and as a result, is the least accurately classified.

Considering inter-subject confusions, we observe that the classifier confuses three subjects the most: Social Sciences, Science, and Miscellaneous. The highest confusion occurs with Miscellaneous questions classified as Social Sciences. This is because Miscellaneous category contains questions similar to Social Sciences but these are too specific to be included in the Social Sciences curriculum or they are so rare that the classifier assigns them to the first subject it finds a clue about. Consider the following miscellaneous question example below. Underlined words are the most important words and we see that the most important words of the miscellaneous question are contained in Social Sciences to which the classifier assigns it. The same logic applies to classify Miscellaneous questions to Science.

Table 8
Individual, ensemble and SERP enhancement accuracies using the bag-of-words model (U-B).

NB	75.9	Term-object search	64.4
HP	55.2	SVM (Linear)	76.8
CNB	76.4	SVM (RB)	66.8
Book search	54.0	Ensemble	78.2
DMNB	77.2	Ensemble + QE	78.4
Object search*	55.7	Ensemble + SERPs	82
Term search*	54.0	Ensemble + SERPs + QE	**83.2**
(*:not included in the ensemble)			

Table 9
Lexical, syntactic and semantic features accuracies.

Lexical*	Syntactic	Semantic	Acc.
WH-WS-QL	P-TU-OS-OSP-WD	S-A-HR-HO-SC-AW-ST-UF-BG	77.62
WH-WS-QL	FP-TU-OS-OSP-WD	S-A-HR-HO-SC-AW-ST-UF-BG	77.32
WH-WS-QL	FP-TU-OS-OSP-WD	S-A-HR-HO-SC-AW	78.78
WH-WS-QL	FP-TU-OS-OSP-WD	HR-HO	78.80
WH-WS-QL	FP-TU-OS-OSP-WD	HR	79.04
WH-WS-QL	FP-TU-OS-OSP	HR	79.22
WH-WS-QL	FP-OS-OSP	HR	77.90
WH-WS-QL	FP-TU-OS	HR	79.24
WS-QL	FP-TU-OS	HR	79.32
QL	FP-TU-OS	HR	79.60
QL	P-TU-OS	HR	79.16
QL	P-OS	HR	78.18
QL	FP-TU-OS	HR	83.30†
-	FP-TU-OS	HR	83.34†
-	TU-OS	HR	**83.58†**
(*:U-B-QE enabled in all)			
(†:SERPs enabled)			
Abbreviations			
WH	Wh-words	S	Synonymy
WS	Word shape	A	Antonymy
QL	Question length	HR	Hypernymy
P	POS tags	HO	Hyponymy
FP	Fine grained POS tags	SC	Side concept
TU	Tagged unigram	AW	Associated with
OS	Object and subject	ST	Similar to
OSP	Object and subject phrase	UF	Used for
U	Unigram	BG	By goal
B	Bigram	WD	Word dependencies
QE	Query expansion	SERP	Search engine result page

Table 10
Confusion matrix for the U-B-TU-OS-HR-QE-SERPS model.

		predicted							
		0	1	2	3	4	5	6	7
actual	(0)Turkish	447	1	0	5	1	31	2	24
	(1)Math.	2	482	0	0	0	15	3	24
	(2)English	9	0	124	2	0	2	1	10
	(3)Rep. Hist.	6	1	0	157	0	31	1	5
	(4)Religion	4	0	0	0	191	17	3	16
	(5)Soc. Sci.	19	7	0	6	2	1279	55	95
	(6)Science	6	18	0	0	0	69	978	54
	(7)Misc.	39	12	1	2	4	161	55	521

Miscellaneous Q: Japonya'da Avukat olmak için ne yapmak lazım?
What should I do to become a lawyer in Japan?
Social Sciences Q.1: Avukatlık mesleğinin toplumdaki yeri ve önemi nedir?
What is the place and the importance of the law profession in the society?
Social Sciences Q.2: Japonya'nın başkenti nedir?
What is the capital of Japan?

Table 11 shows the precision, recall and F1 measurements for each subject for the first model (U-B-QE-SERPS) and the second (U-B-TU-OS-HR-QE-SERPS). English shows the optimal performance (near optimal for the second model) in precision, meaning that there are no false positives for English questions. This may be because the most of the student questions in the English subject are asking for translations of sentences from Turkish to English or homework assignments such as English summaries of known books. Mathematics shows the overall best performance with the highest F1 score. This may be because the questions on this subject are formed as mathematical problems that have direct answers (e.g., If two trains are approaching each other at A and B km/h, when are they going to meet?) and have not much in common with other subjects. It is also the best in terms of recall which denotes the rate of correct identification of the ground truth. Miscellaneous category is the least successful in terms of F1 score. Additionally, we calculated the perplexity of each subject within itself by separating one-tenth of it for the measure. The average perplexities of ten runs are (Social Science: 178.01, Science: 190.30, Miscellaneous:179.63, Religion: 67.75, Turkish: 130.12, English: 81.82, Math :

Table 11
Measurements for each subject.

	U-B-QE-SERPS			U-B-TU-OS-HR-QE-SERPS		
Subject	Precision	Recall	F1	Precision	Recall	F1
Turkish	0.81	0.87	0.84	0.84	0.87	0.86
Math.	0.92	0.92	0.92	0.93	0.92	0.92
English	1.00	0.84	0.91	0.99	0.84	0.91
Rep. Hist.	0.91	0.77	0.84	0.91	0.78	0.84
Religion	0.95	0.83	0.88	0.96	0.83	0.89
Soc. Sci.	0.80	0.88	0.83	0.80	0.87	0.83
Science	0.88	0.86	0.87	0.89	0.87	0.88
Misc.	0.72	0.65	0.68	0.70	0.66	0.67

134.45, Republic History: 106.52). Religion and English seem to have the smallest hence the better language models, while Science, Social Science and Miscellaneous subject categories seem to have the worst.

5. Re-ranking of results for educational queries/questions

In this section, we try to answer two basic questions: 1) Is there a relation between the relevance of query results and match/mismatch of query-document category (i.e., educational subject such as Math and Science) pair? 2) If yes, can we improve the overall relevance by using this category knowledge to make the educational search better?

We use Bing to obtain initially ranked query results to educational questions and to establish a baseline. Note that we may refer to them as questions or queries interchangeably throughout our work since these are the questions that are to be issued as queries. We label these queries with their query results and find out that the answer to the first question is yes (details are given in Section 5.1). After finding query results that may be irrelevant using the classifier described in the previous chapter, we demote them in their lists with different methods. In order to compare these methods, we use the Normalized Discounted Cumulative Gain (NDCG) metric. We perform our evaluation on query sets that vary in query length and being factoid and non-factoid. We further perform significance tests to elaborate on the results.

5.1. Ranking data set and labeling

Because the labeling process of query results is a time-consuming process, we have randomly selected 150 questions from the 5000 labeled question set. Based on a graded relevance scheme, a single annotator have labeled the top-50 results obtained from the Bing search engine for these questions based on grades: {3, 2, 1, 0}.

Grade 3 represents an exact match. Grade 2 represents a good document. Grade 1 represents a document not containing the desired detail but at least having some related concepts. Grade 0 is a totally irrelevant document talking about something else. Our labeling of query results consists of two parts: Labeling for relevance and labeling for the subject. Using these two, we try to learn the relationship between relevance and subject.

We present the relationship between the query subjects and the document subjects in Table 12. If we group relevancy levels 1, 2 and 3 as “relevant” and take relevance 0 as “irrelevant”, it is seen that 99% of the results that do not share the same subject with the question ($s_q \neq s_d$), are found to be irrelevant (i.e., P(irrelevant $\mid$ $s_q \neq s_d$) = 0.99). This means that subject difference may be an indicator of irrelevance in the education context. On the other hand, of the irrelevant pages only 39% found to be from different subjects (i.e., P($s_q \neq s_d \mid$ irrelevant) = 0.39).

5.2. Estimating similarity based on classification

Retrieval systems try to rank documents based on similarity measures that try to estimate $P(d|q)$, the probability of seeing document d given query q, in different ways. We start at the same point. If we take the relevance to be :

$$Sim(d, q) \approx P(d|q) = \frac{P(q|d)P(d)}{P(q)} \tag{6}$$

When we take $P(q|d)$ as $P(s_i|d)P(q|s_i)$ and do this for each s_i:

Table 12
Subject match contingency matrix.

	Relevant	Irrelevant
$s_q = s_d$	4232	1864
$s_q \neq s_d$	9	1192

$$Sim(d, q) \approx \sum_{s_i \in S} \frac{P(s_i|d)P(q|s_i)P(d)}{P(q)} \tag{7}$$

When we take $P(q|s_i)$ as $P(s_i|q)P(q)/P(s_i)$ using Bayes theorem and ignore $P(d)$:

$$Sim(d, q) \approx \sum_{s_i \in S} \frac{P(s_i|d)P(s_i|q)}{P(s_i)} \tag{8}$$

$P(s_i|d)$ and $P(s_i|q)$ are the probabilities of classifying document d and query q in subject s_i, respectively. These are already available using our classification method. $P(c_i)$ can be calculated by dividing the number of instances of c_i in the population by total number of instances: $n_{s_i}/\sum_{s_j \in S} n_{s_j}$. Or, we can further simplify by ignoring this. This reduces the summation to an inner product between probability vectors.

In fact, by using query and document vectors as shown in the example, where each point is represented by a subject probability, we can calculate other vector-based similarity measures. For the sake of this work, we focus on the initial similarity. This similarity will be used as a threshold in the ranking methods below.

6. Proposed ranking methods

In this section, we list the ad hoc methods we use to improve the ranking for educational queries. We use the following notation:
i : the initial rank of a query result, $0 \leq i < 50$
j : the newly calculated rank of a query result, $0 \leq j < 50$
Additionally, the first result was too important that we do not touch the ranking of this document for any query.

6.1. Point-wise

These algorithms try to demote query results that are classified differently than their query. For example, if the query/question is predicted to be in the science subject, then any result classified as other than the science subject will be moved to the below ranks. However, since our classifier, which is described in the previous section is not 100% accurate, we risk demoting pages that can be relevant. These algorithms work basically like this: We find a query result classified differently than the query and its similarity score (described in the previous section) is lower than a threshold (explained earlier), we demote this page to a rank that is calculated by the methods below.

Naive push down If we had a 100% accurate classifier, we would directly demote query results to the end of the list. Since we have 50 pages at most, this method pushes a query result to the end of the list.

Step push down This method is purely based on the current position i of the query result, aiming to provide a smooth reduction mechanism by dividing by the logarithm. The number of ranks the demotion is done increases with the rank (i.e., top results are moved more cautiously).

$$j = i + \left\lceil \frac{i}{\log_2(i+2)} \right\rceil \tag{9}$$

Confusion push down We calculate the probability of confusing one class of queries with others using the confusion matrix from Section 5 (dividing pairwise confusion by total confusion) but this time without the 150 questions we use in the re-ranking experiments. With this method, demotion amount becomes larger when confusion between the subject of the query s_q and the subject of the query result s_i gets smaller. $g(.)$ is a transformation function that scales the probability values which are between 0 and 1 to the desired range. For example, the demoted rank becomes 49 when the confusion is lowest. Questions with types “Social Sciences” and “Miscellaneous” can be confused with each other more often than any other pair. This indicates a higher error rate of classification for such “Social Sciences-Miscellaneous” pairs and motivates us to demote such query results more cautiously. In other words, we trust our classifier less on such classifications and demote accordingly.

$$j = i + \left\lceil \frac{g(P(s_i|c_q))}{\log_2(i+2)} \right\rceil \tag{10}$$

Hybrid push down This method tries to combine Step and Confusion methods. λ is a weighting factor. We take it as 0.5.

$$j = i + \left\lceil \frac{\lambda i + (1-\lambda) g(P(s_i|c_q))}{\log_2(i+2)} \right\rceil \tag{11}$$

6.2. List-wise

Rather than thinking about a pointwise demotion technique, we can calculate a generic score for all results that use both the initial ranking and the classification knowledge.

Linear combination of initial relevance and classification based relevance If we had the original ranking scores of query

results, we could have used them. Since we do not, for estimating the initial relevance, we use a logarithmic relevance function in the form $-a\log(i+1)+b$ that starts with high relevance at the first few ranks and decreases quickly but the rate of decrease diminishes as it gets to later ranks, which is similar to Normalized Discounted Cumulative Gain (NDCG) metric described in Section 7.1. a and b are taken as 0.4 and 2.8, respectively. These values give us a smooth function between 2.8 and 1.2 for i = 0 to i = 49. This way, we simulate the decreasing relevance with increasing rank. The idea is that the expected relevance of the first results is not supposed to be 3 (i.e., perfect) but close to that (2.8) and the expected relevance at rank 49 is around 1, not 0 since the baseline (the initial ranking by Bing) also has a good ranking.

Then, we incorporate this initial relevance estimation and the classification knowledge in Eq. 12 by using λ as a tuning parameter between 0 and 1. In our experiments, we set it as 0.7.

$$rel(q, d_i) = \lambda rel_{init}(i) + (1-\lambda)Sim(q, d_i) \tag{12}$$

7. Ranking evaluation

In this section, we provide the evaluation results of our methods that try to improve educational query result ranking based on classification.

7.1. Methodology and metric

Using the set of query results with relevance judgments, we would be able to evaluate the performance of the proposed methods. In information retrieval, for binary judgments, many evaluation metrics are used including Precision, Recall, Mean Average Precision, and Precision@k. Since we employ graded relevance, we use the highly common NDCG metric to evaluate our methods at various result ranks. NDCG is the normalized version of Discounted Cumulative Gain (DCG) which gives more importance to higher ranked documents. In DCG (see Eq. (13)), we sum the relevance of documents with a discount according to their rank. This forces us to rank more relevant documents in lower ranks.

Ideal DCG (IDCG) is the ideal version of ranking; sorted by relevance values. We get NDCG by dividing the DCG with IDCG, which is a metric of how close we are to the ideal ranking (see Eq. (14)).

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{\log_2 i+1} \tag{13}$$

$$NDCG@k = \frac{DCG@k}{IDCG@k} \tag{14}$$

We choose the default ranking of Bing as our baseline (listed as the *default* method in the graphs). Note that this is a strong baseline that employs various signals to generate query rankings. We mostly pay attention to top 10 query results but provide significance tests for ranks up to 50.

7.2. Overall performance

We run our methods on the set of 150 queries. NCDG@k results (up to @10) are given in Fig. 4. For earlier ranks, the Naive method is above the baseline but starts to decline for not being able to compensate for FP errors (i.e., demoting relevant pages).

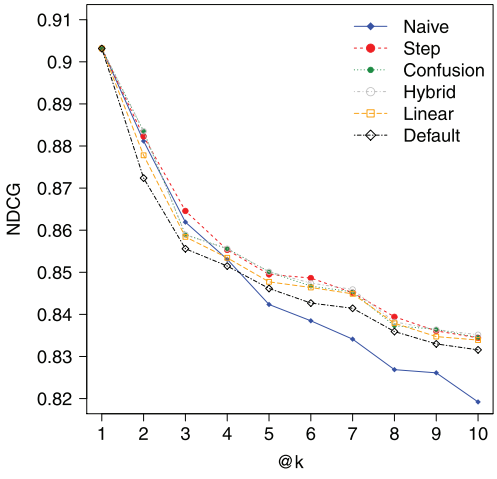

Fig. 4. NDCG Comparison on the Whole Set.

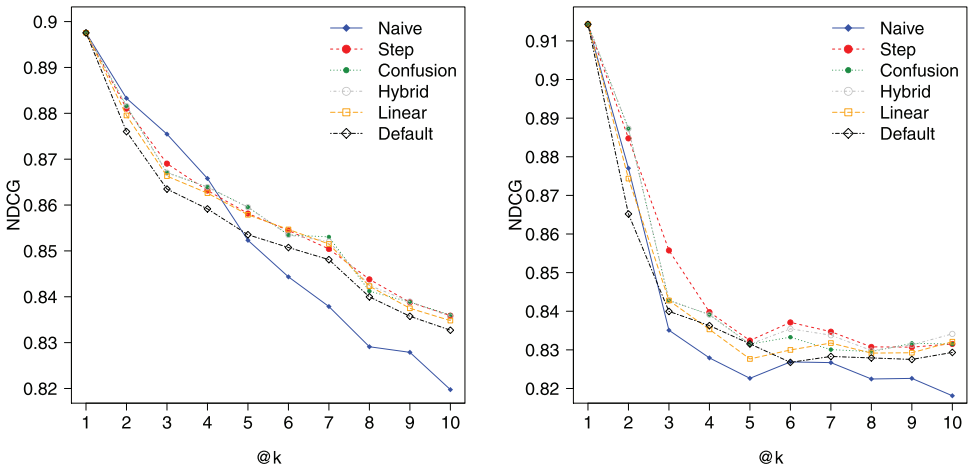

Fig. 5. Changing query length NDCG comparison.

Although different in nature, all the other methods seem to outperform the baseline. The Step method seems to work despite the fact that it only depends on the initial rank of the query results.

7.3. Query length

In informational retrieval systems, some studies suggest that query length can be associated with retrieval performance since they take part in the smoothing processes of retrieval systems and longer queries up to some point contain more knowledge about user intent in earlier (Zhai & Lafferty, 2001) and later (Azzopardi, 2015) studies. We also want to briefly explore the effect of query length for our query set which consists of educational questions that are in natural language form. Our query set consists of 100 queries that have 6 or fewer tokens and 50 queries with more than 6 tokens. 6 is the median length of the queries in our entire set described in the previous sections, meaning that our choice of 6 is not arbitrary. In Fig. 5a, we see the retrieval effectiveness of all methods for shorter queries. Although this looks similar to the effectiveness performance on the whole set in Fig. 4, we see that the Naive method performs slightly different at earlier ranks. Interestingly, the Naive method achieves the best NDCG for top2, top3, and top4 ranks. For long queries, in Fig. 5b, the Naive method is not as much separated from the other methods as it is in other evaluations. We associate this with the assumption that longer queries are better classified than shorter ones or they return better results. The latter part of the assumption goes parallel only with the first 2 ranks. Compared to the left plot, NDCG values drop significantly after rank 2. Our methods seemingly perform better in longer queries with respect to the baseline, as they do for short queries. In long queries, the methods perform around 2% better than the baseline, whereas in short queries the performance gain is less than 2%.

7.4. Factoid vs non-Factoid questions

In Question Answering research, factoid and non-factoid questions attracted a wide span of attention such as Bian, Liu, Agichtein, and Zha (2008) and Soricut and Brill (2004). Factoid questions are the ones with simple and mostly one word or phrase answers (e.g., When did the Wright brothers fly their first airplane?).

A non-factoid question, on the other hand, can be very hard. This type of questions can be in many forms such as opinion seeking, how, why (e.g., What are the benefits of ancient civilizations that existed in Anatolia and Mesopotamia?).

From our 150 question dataset, we labeled the first 100 questions on being factoid or non-factoid. We found 73% to be factoid and 27% to be non-factoid. We run our methods on these questions. Fig. 6a shows that relevancy is very high with factoid type queries and it does not diminish as it is in the other types of queries evaluated in this section. We also notice that the Linear method outperforms the others only in this type of queries. For non-factoid queries in Fig. 6b, relevancy declines fast. Tu et al. (2008) report a similar result concerning “open and close-ended” educational questions where the results for close-ended questions were more relevant than those of open-ended ones.

7.5. Significance tests

In order to show that our results are significant, we use the paired *t*-test at each rank comparing the resulting rankings of each method to the default ranking. We list the calculated p values at each rank for every method compared to the baseline in Table 13. Note that even though we have a small sample size, we observe significant results. Confusion and Hybrid methods seem to have p values smaller than 0.05, and even smaller than 0.001 most of the time. The Hybrid method achieves significantly better results

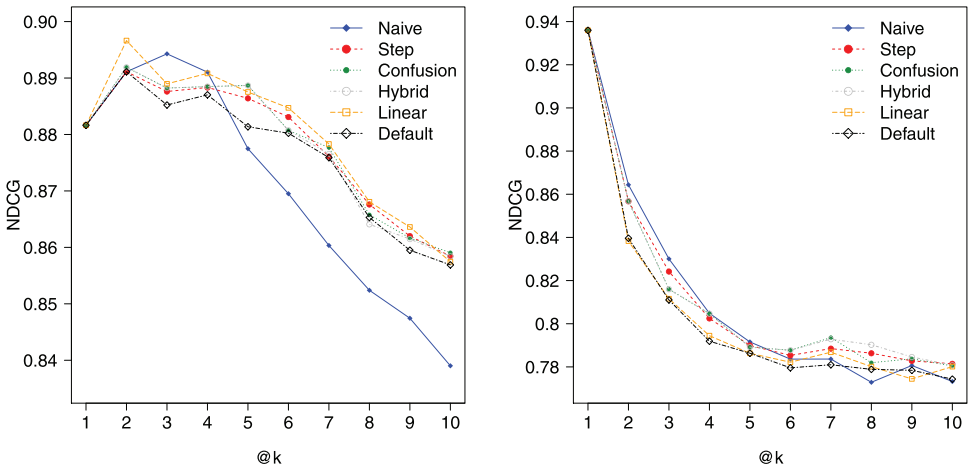

Fig. 6. Factoid and non-factoid NDCG comparison.

Table 13
Significance tests p values.

Method	NDCG													
	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10	@20	@30	@40	@50
Naive	=	0.11	0.18	0.39	-	-	-	-	-	-	-	-	-	-
Step	=	0.06	**0.03**	0.01	0.07	0.008	0.04	**0.06**	0.07	0.03	0.04	0.0008	0.03	0.01
Confusion	=	**0.03**	**0.03**	**0.005**	**0.03**	0.003	**0.007**	0.15	**0.0008**	0.002	**0.001**	0.0006	**0.009**	0.002
Hybrid	=	**0.03**	**0.03**	**0.005**	**0.03**	**0.001**	0.02	0.10	0.02	**0.0003**	0.05	**0.00005**	0.02	**0.0009**
Linear	=	0.16	0.20	0.14	0.29	0.07	0.08	0.23	0.26	0.17	0.27	0.007	0.11	0.09

consistently up to rank 6. We see that the significance increases as the rank becomes higher since the number of comparison points also increases.

8. Conclusion

We showed that the query and the query result subjects matter while using web search for educational queries. We first developed an ensemble educational query classifier using lexical, syntactic and semantic features with 83.5% accuracy. Using this classifier, we provided methods for improving the ranking of query results in the education context. We used a general purpose search engine as the baseline and improved the ranking of query results returned from this system. We employed the NDCG metric to evaluate our methods with respect to the baseline on different types of queries. Our results show significant relevancy improvement for educational queries.

After showing that the ad-hoc methods work for our purpose, as a future work direction, we can consider a more generic learning to rank approach. For such an approach, possible learning features include the current ranking position of the result, an estimation of the initial relevance, query-document classification similarity (the one we described in Section 5), cosine, Euclidean and correlation distances derived from query and result subject probability vectors. Of course, a learning to rank approach may require additional labeling of query results.

Declaration of interest

None

Acknowledgment

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) [grant number 113E065]

References

Azzopardi, L. (2015). *Theory of retrieval: The retrievability of information. Proceedings of the 2015 international conference on the theory of information retrieval*. ACM3–6.

Bian, J., Liu, Y., Agichtein, E., & Zha, H. (2008). *Finding the right facts in the crowd: Factoid question answering over social media. Proceedings of the 17th international conference on world wide web*. ACM467–476.
Bilal, D. (2002). Children's use of the yahooligans! web search engine. iii. cognitive and physical behaviors on fully self-generated search tasks. *The Journal of the Association for Information Science and Technology, 53*(13), 1170–1183. https://doi.org/10.1002/asi.10145.
Bilgin, O., Çetinoğlu, Ö., & Oflazer, K. (2004). Building a wordnet for turkish. *Romanian Journal of Information Science and Technology, 7*(1–2), 163–172.
Bloom, B. S., & Engelhart, M. D. (1969). *Taxonomy of educational objectives: The classification of educational goals: By a committee of college and university examiners: Handbook 1*. David McKay.
Brainly Inc. (2018). Brainly - career & press. http://brainly.co. Accessed: 2018-08-07.
Buchholz, S., & Marsi, E. (2006). *Conll-x shared task on multilingual dependency parsing. Proceedings of the tenth conference on computational natural language learning*. Association for Computational Linguistics149–164.
Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H. C., & Vursavas, O. M. (2008). Information retrieval on turkish texts. *The Journal of the Association for Information Science and Technology, 59*(3), 407–421. https://doi.org/10.1002/asi.v59:3.
Cao, X., Cong, G., Cui, B., Jensen, C. S., & Zhang, C. (2009). *The use of categorization information in language models for question retrieval. Proceedings of the 18th acm conference on information and knowledge managementCIKM '09*New York, NY, USA: ACM265–274. https://doi.org/10.1145/1645953.1645989.
Chau, M., & Chen, H. (2003). Comparison of three vertical search spiders. *Computer, 36*(5), 56–62.
Chen, L. (2014). *Understanding and exploiting user intent in community question answering* Ph.D. thesis.
Compete Site Analytics (2015). answers.yahoo.com UVs for May 2015 | Compete.https://siteanalytics.compete.com/answers.yahoo.com/. Accessed: 2015-05-04.
Dersimiz.com (2018). Büyük Terimler Sözlüğü Arşivi, Ders Terimleri. http://www.dersimiz.com/terimler-sozlugu/. Accessed: 2018-08-07.
Diligenti, M., Gori, M., & Maggini, M. (2002). *Web page scoring systems for horizontal and vertical search. Proceedings of the 11th international conference on world wide web*. ACM508–516.
EOdev (2018). Eodev.com - Ödevlerin yeni boyutu. http://eodev.com. Accessed: 2018-08-07.
Eryigit, G. (2014). *ITU Turkish nlp web service. 14th conference of the european chapter of the association for computational linguistics*1–4.
Eryiğit, G., Nivre, J., & Oflazer, K. (2008). Dependency parsing of turkish. *Computational Linguistics, 34*(3), 357–389.
Espina, A., & Figueroa, A. (2017). Why was this asked? automatically recognizing multiple motivations behind community question-answering questions. *Expert Systems with Applications, 80*, 126–135.
Eustace, J., Wang, X., & Li, J. (2014). Approximating web communities using subspace decomposition. *Know.-Based Syst. 70*(C), 118–127. https://doi.org/10.1016/j.knosys.2014.06.017.
Figueroa, A., & Neumann, G. (2016). Context-aware semantic classification of search queries for browsing community question?answering archives. *Knowledge-Based Systems, 96*, 1–13. https://doi.org/10.1016/j.knosys.2016.01.008.
Gazan, R. (2011). Social q&a. *Journal of the American Society for Information Science and Technology, 62*(12), 2301–2312.
Ghosh, A., & Kleinberg, J. (2013). *Incentivizing participation in online forums for education. Proceedings of the fourteenth acm conference on electronic commerce*. ACM525–542.
Gurevych, I., Bernhard, D., Ignatova, K., & Toprak, C. (2009). *Educational question answering based on social media content. Aied*133–140.
Haris, S. S., & Omar, N. (2012). *A rule-based approach in bloom's taxonomy question classification through natural language processing. Computing and convergence technology (iccct), 2012 7th international conference on*. IEEE410–414.
Harper, F., Weinberg, J., Logie, J., & Konstan, J. (2010). Question types in social q&a sites. *First Monday, 15*(7).
Harper, F. M., Moy, D., & Konstan, J. A. (2009). *Facts or friends?: Distinguishing informational and conversational questions in social q&a sites. Proceedings of the sigchi conference on human factors in computing systems*. ACM759–768.
Harper, F. M., Raban, D., Rafaeli, S., & Konstan, J. A. (2008). *Predictors of answer quality in online q&a sites. Proceedings of the sigchi conference on human factors in computing systemsCHI '08*New York, NY, USA: ACM865–874. https://doi.org/10.1145/1357054.1357191.
Hermjakob, U. (2001). *Parsing and question classification for question answering. Proceedings of the workshop on open-domain question answering-volume 12*. Association for Computational Linguistics1–6.
Huang, Z., Thint, M., & Qin, Z. (2008). *Question classification using head words and their hypernyms. Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics927–936.
ITU Kemik (2018). Veri kümelerimiz. http://www.kemik.yildiz.edu.tr/?id=28. Accessed: 2018-08-07.
Jain, A., & Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(2), 153–158. https://doi.org/10.1109/34.574797.
Kammerer, Y., & Bohnacker, M. (2012). *Children's web search with google: The effectiveness of natural language queries. Proceedings of the 11th international conference on interaction design and childrenIDC '12*New York, NY, USA: ACM184–187. https://doi.org/10.1145/2307096.2307121.
Li, B., Jin, T., Lyu, M. R., King, I., & Mak, B. (2012). *Analyzing and predicting question quality in community question answering services. Proceedings of the 21st international conference on world wide web*. ACM775–782.
Li, H., Samei, B., Olney, A. M., Graesser, A. C., & Shaffer, D. W. (2014). *Question classification in an epistemic game. 3rd workshop on intelligent support for learning in groups (islg) at the 12th international conference on intelligent tutoring systems*. Springer.
Li, X., & Roth, D. (2002). *Learning question classifiers. Proceedings of the 19th international conference on computational linguistics-volume 1*. Association for Computational Linguistics1–7.
Li, X., & Roth, D. (2006). Learning question classifiers: The role of semantic information. *Natural Language Engineering, 12*(03), 229–249.
Liu, J., Shen, H., & Yu, L. (2017). Question quality analysis and prediction in community question answering services with coupled mutual reinforcement. *IEEE Transactions on Services Computing, 10*(2), 286–301.
Loni, B. (2011a). Enhanced question classification with optimal combination of features. Master's thesis. Delft University of Technology.
Loni, B. (2011b). *A survey of state-of-the-art methods on question classificationTechnical Report*. Delft University of Technology.
Mao, J. (2014). Social media for learning: A mixed methods study on high school students' technology affordances and perspectives. *Computers in Human Behavior, 33*, 213–223.
Metzler, D., & Croft, W. B. (2005). Analysis of statistical question classification for fact-based questions. *Information Retrieval, 8*(3), 481–504.
Microsoft Azure (2018). Bing web Search API | Microsoft Azure. https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/. Accessed: 2018-08-07.
Mishra, M., Mishra, V. K., & Sharma, H. (2013). Question classification using semantic, syntactic and lexical features. *International Journal of Web & Semantic Technology, 4*(3), 39.
Msxlabs (2018). Okula Destek. http://msxlabs.com. Accessed: 2018-08-07.
Ochoa, X., & Duval, E. (2008). Relevance ranking metrics for learning objects. *IEEE Transactions on Learning Technologies, 1*(1), 34–48.
Oflazer, K. (2014). Turkish and its challenges for language processing. *Language Resources And Evaluation, 48*(4), 639–653. https://doi.org/10.1007/s10579-014-9267-2.
Pal, A., & Konstan, J. A. (2010). *Expert identification in community question answering: Exploring question selection bias. Proceedings of the 19th acm international conference on information and knowledge managementCIKM '10*New York, NY, USA: ACM1505–1508. https://doi.org/10.1145/1871437.1871658.
Palomera, D., & Figueroa, A. (2017). Leveraging linguistic traits and semi-supervised learning to single out informational content across how-to community question-answering archives. *Information Sciences, 381*, 20–32.
Purcell, K., Rainie, L., Heaps, A., Buchanan, J., Friedrich, L., Jacklin, A., et al. (2012). How teens do research in the digital world. *Pew Internet & American Life Project*.
Quora (2018). Quora - the best answer to any question. https://www.quora.com/. Accessed: 2018-08-07.
Rieh, S. Y., Collins-Thompson, K., Hansen, P., & Lee, H.-J. (2016). Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science, 42*(1), 19–34.

Sangodiah, A., Muniandy, M., & Heng, L. E. (2015). Question classification using statistical approach: A complete review. *Journal of Theoretical and Applied Information Technology, 71*(3).
Schacter, J., Chung, G. K. W. K., & Dorr, A. (1998). Children'S internet searching on complex problems: performance and process analyses. *Journal Of The American Society For Information Science And Technology, 49*(9), 840–849. https://doi.org/10.1002/(SICI)1097-4571(199807)49:9<840::AID-ASI9>3.3.CO;2-4.
Silva, J., Coheur, L., Mendes, A. C., & Wichert, A. (2011). From symbolic to sub-symbolic information in question classification. *Artificial Intelligence Review, 35*(2), 137–154.
Soricut, R., & Brill, E. (2004). *Automatic question answering: Beyond the factoid. Hlt-naacl*57–64.
Stack Exchange (2018). Stack exchange: Hot questions. http://stackexchange.com/. Accessed: 2018-08-07.
Stack Overflow (2018). Stack Overflow. http://www.stackoverflow.com. Accessed: 2018-08-07.
Tu, Y.-W., Shih, M., & Tsai, C.-C. (2008). Eighth graders' web searching strategies and outcomes: The role of task types, web experiences and epistemological beliefs. *Computers & Education, 51*(3), 1142–1153.
Türk Dİl Kurumu (2018). Büyük Türkçe Sözlük - Türk Dİl Kurumu. http://tdk.gov.tr/index.php?option=com_bts&view=bts. Accessed: 2018-08-07.
Usta, A. (2015). Optimization of an educational search engine using learning to rank algorithms. Master's thesis. Bilkent University.
Usta, A., Altingovde, I. S., Vidinli, İ. B., Ozcan, R., & Ulusoy, Ö. (2014). *How k-12 students search for learning?: Analysis of an educational search engine log. Proceedings of the 37th international acm sigir conference on research & development in information retrieval.* ACM1151–1154.
Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., & Duval, E. (2012). Context-aware recommender systems for learning: A survey and future challenges. *IEEE Transactions on Learning Technologies, 5*(4), 318–335. https://doi.org/10.1109/TLT.2012.11.
Vikisözlük (2018). Vikisözlük: Özgür Sözlük. https://tr.wiktionary.org/. Accessed:2018-08-07.
Vitamin Eğitim (2018). Vitamin Eğitim. http://www.vitaminegitim.com. Accessed: 2018-08-07.
Vlasák, B. M. (2015). Online school–educational content classification and recommendation. Master's thesis. Masaryk University.
Wiktionary (2018). Wiktionary, the free dictionary. https://en.wiktionary.org. Accessed: 2018-08-07.
Wu, Y., Hori, C., Kashioka, H., & Kawai, H. (2015). Leveraging social q&a collections for improving complex question answering. *Computer Speech & Language, 29*(1), 1–19.
Yahoo! Answers (2018). Yahoo! answers education & reference category. https://answers.yahoo.com/dir/index?sid=396545015. Accessed: 2018-08-07.
Yahya, A. A., & Osman, A. (2011). *Automatic classification of questions into Bloom's cognitive levels using support vector machines. The international arab conference on information technology*1–6.
Yang, J., Tao, K., Bozzon, A., & Houben, G.-J. (2014). *Sparrows and owls: Characterisation of expert behaviour in stackoverflow. User modeling, adaptation, and personalization.* Springer266–277.
Yusof, N., & Hui, C. J. (2010). *Determination of bloom's cognitive level of question items using artificial neural network. Intelligent systems design and applications (isda), 2010 10th international conference on.* IEEE866–870.
Zhai, C., & Lafferty, J. (2001). *A study of smoothing methods for language models applied to ad hoc information retrieval. Proceedings of the 24th annual international acm sigir conference on research and development in information retrieval.* ACM334–342.
Zhang, D., & Lee, W. S. (2003). *Question classification using support vector machines. Proceedings of the 26th annual international acm sigir conference on research and development in informaion retrieval.* ACM26–32.