Natural Language Engineering **24** (5): 725–762. © Cambridge University Press 2018
doi:10.1017/S1351324918000141

*A closed-domain question answering framework using reliable resources to assist students**

CANER DERİCİ[1], YİĞİT AYDIN[2],
ÇİĞDEM YENİALACA[2],
NİHAL YAĞMUR AYDIN[1],
GÜNİZİ KARTAL[2], ARZUCAN ÖZGÜR[1], and
TUNGA GÜNGÖR[1]

[1]*Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey*
e-mail: caner.derici@boun.edu.tr, yagmur.aydin@boun.edu.tr, arzucan.ozgur@boun.edu.tr, gungort@boun.edu.tr

[2]*Department of Computer Education and Educational Technology, Boğaziçi University, Istanbul, Turkey*
e-mail: yigitaydinn@gmail.com, yenialacacigdem@gmail.com, gunizi.kartal@boun.edu.tr

(*Received 14 November 2016; revised 12 March 2018; accepted 14 March 2018; first published online 10 April 2018*)

Abstract

This paper describes a question answering framework that can answer student questions given in natural language. We suggest a methodology that makes use of reliable resources only, provides the answer in the form of a multi-document summary for both factoid and open-ended questions, and produces an answer also from foreign resources by translating into the native language. The resources are compiled using a question database in the selected domains based on reliability and coverage metrics. A question is parsed using a dependency parser, important parts are extracted by rule-based and statistical methods, the question is converted into a representation, and a query is built. Documents relevant to the query are retrieved from the set of resources. The documents are summarized and the answers to the question together with other relevant information about the topic of the question are shown to the user. A summary answer from the foreign resources is also built by the translation of the input question and the retrieved documents. The proposed approach was applied to the Turkish language and it was tested with several experiments and a pilot study. The experiments have shown that the summaries returned include the answer for about 50–60 percent of the questions. The data bank built for factoid and open-ended questions in the two domains covered is made publicly available.

1 Introduction

The number of information sources and the amount of information that exist on electronic environments and the World Wide Web are increasing steadily. Users

*This work was supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK) under the grant number 113E036. We would like to thank Çağıl Uluşahin Sönmez for her contribution in the Google Translate interface of the research.

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

cater for their information needs related to their topics of interest on these resources using several technologies, including search engines. However, such technologies have two main deficiencies due to being general purpose, not tailored to a particular domain, and making use of natural language processing (NLP) methods in a very restricted manner. The first one is not being able to take into account the natural language structures related to the domain and thus especially failing to satisfy complex information needs. The second one is the unreliable nature of the resources (Web documents) used and the inconsistent results obtained.

Question answering (QA) is the task of answering questions expressed by the user in natural language. Given a natural language question, the first step performed by a QA system is the analysis of the question for mainly two reasons. The first one is the extraction of the content that is related to what is actually asked for. The question may contain irrelevant details and the same content may be expressed in different ways. Since the answer to the question is to be searched within the documents in the collection, it should be converted to a representation suitable for matching with the contents of the documents. The second reason is the identification of the type of the question, which is important for returning a sensible answer.

In the case of QA on unstructured natural language documents, an information retrieval (IR) system is used. A query is built from the question that is analyzed and represented in a suitable form. The IR system returns a set of documents relevant to the query. The documents are then processed and the answer is extracted by taking the question type into account. We can divide the questions into two broad groups: factoid questions and open-ended questions. A factoid question is defined as a question whose answer is formed of one or a few words, and that has a unique answer in all the resources. An open-ended question is one whose answer is a descriptional text rather than a single word or phrase, and the answer may be different in different resources. Though the tasks of IR and QA seem related, there is a basic difference. While the former inputs a keyword-based query and returns documents as output, the latter one accepts a question in natural language form and returns just the answer.

In this study, we aim at developing a QA framework that meets the information needs (questions) of students using reliable resources. We name the system as *HazırCevap*, which is an idiom in Turkish that denotes a person who can answer questions easily. The system enables the students to be able to use resources in foreign languages, and presents the results in the form of a combined and coherent summary. The proposed approach consists of the phases of asking questions by the students in natural language (Turkish), analysis of the questions using natural language processing techniques, identifying Turkish and English resources to answer the questions, translating the foreign resources to Turkish, and analysis, combination and summarization of the related resources oriented toward the requested information need. We aim at making available the resources that are out of the scope for the students due to the language barrier and enabling the students' access to question-based, accurate and reliable information using an environment specialized for answering the students' information needs.

The approach proposed in this work is novel with respect to its subject and methodology. Each part of the framework built is a research topic by itself and is

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

subject to a great amount of research for well-studied languages. However, there is no study that aims at combining these parts as proposed in the current framework. Given a user question, the answers are derived from both the native language and foreign language resources and the final answer is offered in the form of a single summary in the user's native language. Using summaries as answers to questions aims at presenting the user additional information related to the context of the question. In addition, by being oriented to Turkish educational programs, comprising components specialized for topics covered by those programmes, and having a very limited number of research and development activities on these subjects focused on the Turkish language contributes to the novelty of the study.

One of the main contributions of the study is the data bank of open-ended and factoid questions in the biology and geography domains. Questions in the corpus were annotated with their answers as well as their focus, modifier, and class information. We also introduce a novel representation model for question formulation making use of these semantic features extracted from the questions. Multi-lingual resources are used to utilize the rich information source of the web resources in English.

The research objective in this study is building a framework that communicates with students in natural language and provides accurate and explanatory answers in response to their questions. Question analysis and question representation phases are tailored to the characteristics of student questions by designing suitable question classes and adjusting term weights. In addition to the components of question analysis, document retrieval, and answer extraction that exist in traditional QA systems, the modules of resource compilation, translation, and summarization are incorporated in the framework. They contribute to increasing the reliability of the answers and make it possible to use foreign resources and to output detailed answers. We test the plausibility of using such a comprehensive QA framework in educational environments.

The rest of the paper is organized as follows. Section 2 gives an overview of previous work. Section 3 shows the general architecture of the system. The following three sections are devoted to the components in the system. Section 4 explains how the resources used in the system were determined and compiled. The question analysis module and the question representation model are explained in Section 5. This is followed by the answer generation module in Section 6. Section 7 shows the experiments and the pilot study. Section 8 concludes the paper.

2 Related works

2.1 Question analysis and question answering

The topic of QA was covered from different perspectives in several survey studies (Kolomiyets and Moens 2011; Bouziane et al. 2015; Mishra and Jain 2016; Höffner et al. 2016; Diefenbach et al. 2017; Utomo, Suryana and Azmi 2017). Kolomiyets and Moens (2011) give a list of the approaches used in QA. The approaches range from the simplest ones that represent the question and the documents using bag of words or some morpho-syntactic forms of the words to the advanced models

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

that extract the discourse relations within the sentences. Diefenbach et al. (2017) divide the QA process into five steps: question analysis (extracting named entities, identifying dependencies between words), phrase mapping (mapping parts of a question with entities in a knowledge base using string and semantic similarities), disambiguation (solving segmentation and mapping ambiguities), query construction (converting the question into a query by using templates, semantic parsing, or machine learning approaches), and querying distributed knowledge (answering questions using multiple knowledge sources). They analyze 27 QA systems with respect to these criteria.

Mishra and Jain (2016) give another classification of QA methods, where the approaches are divided into four broad categories. These categories correspond to systems that are based on data mining, IR, natural language understanding, and knowledge retrieval and discovery techniques. The methods used in each category and other aspects are discussed. The QA field is also analyzed from other perspectives such as the question types, whether the domain is open or restricted, and the types and properties of resources. Höffner et al. (2016) focus on QA on the semantic web. They identify seven challenges that should be addressed by semantic QA systems and analyze sixty-two systems with respect to these challenges. The challenges are lexical gaps between words, ambiguity, multilingualism in web documents, complex questions (e.g., nested questions), distributed knowledge, procedural, temporal, and spatial questions (in addition to factoid, list, and yes/no questions), and mapping questions to templates. They observe that the issues of lexical gaps and ambiguities are handled by most of the systems, while the other issues are dealt with only in a few systems.

A pioneering attempt in the QA field was initiated by Text Retrieval Conference (TREC)[1] in 1999. The goal was generating short answers to factoid and list questions that can be drawn from any domain (Olvera-Lobo and Gutierrez-Artacho 2015). The tasks included in the track have evolved in time. The problem was first defined as, given a question, returning a text string consisting of a complete answer. The scope of the track was then extended to include tasks such as the passages task (returning a short passage) and the complex interactive task (addressing information needs more complex than factoid questions). The track ended in 2007. It has been restarted in 2015 under the name of LiveQA, which targets answering real user questions in real time. We can cite CLEF,[2] NTCIR,[3] and BioASQ[4] as other shared tasks related to the QA problem.

The early systems AnswerBus (Zheng 2002), AskMSR (Brill, Dumais and Banko 2002), and START (Katz 1997) can be regarded as the leading studies in the field. AnswerBus is an open-domain system that uses general purpose search engines such as Google and Yahoo to find the best matching web pages for the given question. It uses a bag-of-words approach and employs several search engines. The main

[1] http://trec.nist.gov
[2] http://nlp.uned.es/clef-qa
[3] http://research.nii.ac.jp/qalab/task.html
[4] http://www.bioasq.org

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

drawback of AnswerBus is that it can only return a web page that is most likely to include an answer to the question, but does not return an actual answer. On the other hand, AskMSR uses question analysis and answer generation techniques rather than a bag-of-words approach. It generates several possible rewrites of the given question using simple morphological variations. Using these different representations, the system forms different queries. Candidate answers to the queries are generated using n-gram frequencies. The START system uses large numbers of knowledge sources to build a knowledge-based representation and it can answer open-ended and factoid questions by matching on structured data.

It is mostly believed that identifying the question class before attempting to generate an answer helps in increasing the relevance of the answer. Wu et al. (2015) argue that a question contains some cue expressions depending on its question class. They propose a method that learns such cue expressions for each question type from social question–answer collections. The proposed question type-specific method was also compared with question-specific (Chen, Zhou and Wang 2006) and monolingual translation-based (Bernhard and Gurevych 2009) approaches. Figueroa and Neumann (2016) propose a method for classifying question-like search queries into a set of twenty-six pre-defined semantic categories. For classification, a maximum entropy model is used and the training set is formed of the sequences of queries in the search sessions of the users. Pechsiri and Piriyakul (2016) target two types of open-ended questions (“why” questions and “how” questions). Question class identification is performed by employing a machine learning algorithm on question patterns. Answer extraction is then done by returning the starting and ending points of the so-called elementary discourse units, which are similar to sentences or clauses. The answer extraction accuracy was reported around ninety percent by manual evaluation.

Identification of suitable features for question classification or answer generation is a critical task in QA systems. Figueroa and Neumann (2016) extract eleven different features from the queries, which are surface, syntactic (such as lexical chains), and semantic (such as named entities) features. The features are used to infer the semantic class of the question. Khodadi and Abadeh (2016) use genetic programming in order to obtain new features by combining the elementary features. A set of features based on lexical, syntactic, and semantic properties of the sentences is determined. The paragraphs and sentences are ranked based on the feature set for definitional and factoid questions. Another work proposes a method based on reinforcement learning that generates a multi-document summary as answer to a question (Chali, Hasan and Mojahid 2015). They represent the sentences in terms of static features, which are features used in the text summarization context, and dynamic features. Yang et al. (2015a) aim at converting a natural language question into a logical form using the Freebase database as the knowledge source. A question is represented in terms of lexical features (words) and logical features, which are the category of the question, topic, entity type, and a predicate about the answer sought. The words are linked to logical features and an answer is generated using the similarities of question words and logical features of potential answers. They obtained thirty-seven percent precision and fifty-six percent recall (forty-five percent F-measure) on a dataset of about 2,000 question–answer pairs.

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

There are some systems that have been designed to answer questions asked in well-known games. The most famous one is IBM Watson, which competes on the game show known as Jeopardy!.[5] Since it has been especially designed to answer questions on this show, a major amount of the system has focused on the type of questions used in the game. However, the NLP and machine learning modules and the overall architecture of the system is designed to work as an open domain QA system (Fan et al. 2012). The QA pipeline consists of the stages of question analysis and parsing (Lally et al. 2012; McCord, Murdock and Boguraev 2012), document retrieval (Chu-Carroll et al. 2012b), candidate answer generation (Chu-Carroll et al. 2012a; Murdock et al. 2012b), scoring candidate answers (Murdock et al. 2012a), and merging and ranking the answers (Gondek et al. 2012). Another game-oriented QA system is the system that was developed for the popular game "who wants to be a millionaire?" (Molino et al. 2015). The system makes use of Wikipedia and DBpedia as external knowledge sources. A given question is processed using a pipeline of natural language modules. The query built is input to three different search engines and the related passages are extracted. These passages are processed by a set of filters that rank them and select the best one. Then, a number of scoring criteria are applied to the candidate passages to extract the answer.

Some of the works in the literature analyze different stages in a QA framework. Habibi, Mahdabi and Popescu-Belis (2016) focus on query expansion for a conversational environment. They extract the question context using a conversation fragment and identify the important keywords in the context with a topic similarity metric. The keywords are also refined using Wikipedia and WordNet as external resources and by learning word embeddings of similar words. Bordes, Chopra and Weston (2014) and Yih, He and Meek (2014) also use word embeddings for open domain QA. Momtazi and Klakow (2015) propose models for selecting the most relevant answer sentences. They argue that the classical document retrieval methods are not suitable for sentence retrieval due to the data sparsity problem and the importance of exact matching. They employ two language modeling approaches, which are class-based and trigger language models. Shekarpour et al. (2015) focus on transforming a question into a set of segments in order to search on fragmented data in different resources. The segments are obtained by first identifying the important keywords in the question and then combining these keywords into meaningful phrases. The best segmentation is determined using a query graph and a hidden Markov model setting.

As in other fields in NLP, deep learning architectures gained popularity in the QA domain. The goal is alleviating the problem of feature engineering and complex linguistic processing. The methods designed for this purpose encode the semantic knowledge within the questions and the documents, and generate answers based on these encodings. Yu et al. (2014) formulate the problem as a binary classification problem: given a question, a sentence in a document is either an answer sentence or not. They employ a convolutional neural network (CNN) model on bigram word features. The experiments on a set of factoid questions compiled from TREC data

[5] http://www.jeopardy.com.

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

yielded around 0.71 MAP (mean average precision) and 0.78 MRR (mean reciprocal rank) scores. Yang, Yih and Meek (2015b) introduced the WikiQA dataset, which is used as a benchmark in the field, and implemented and compared several methods on this dataset. They observed that combining the CNN architecture with features based on common word counts in the question and answer sentence shows the best performance.

Long short-term memory (LSTM) models are frequently used for sequence modeling in NLP. Wang and Nyberg (2015) employ bidirectional LSTM (biLSTM) networks for answer sentence selection. The questions and answers are converted into distributional representations and matching between a question and candidate answers is done using a similarity metric. Iyyer et al. (2014) use a dependency tree recursive neural network to learn question and answer representations jointly. Each word is given to the network together with its dependent words and dependency relations. The model was tested in two domains and compared with several baselines. Bordes, Weston and Usunier (2014) propose a method for QA on knowledge bases. They convert questions and knowledge base triplets into vector embeddings. A scoring function adapted from an image labeling work is used to calculate the similarity of a question with candidate triplets. The experiments with WikiAnswers questions on the ReVerb knowledge base showed around 0.60–0.73 F-measure performance. Some other works that use deep learning frameworks are given by Dong et al. (2015) that uses a multi-column CNN, Xiong, Merity and Socher (2016) that uses a dynamic memory network, and Feng et al. (2015) that employs several CNN architectures.

There are a few QA works for the Turkish language. İlhan et al. (2008) find the best fitting answer from a pre-formed answers database using data mining techniques. The question is converted into vector space model and the answer is selected using a similarity metric. In another study on Turkish, the authors make use of surface level patterns called answer patterns (Er and Çiçekli 2013). They first extract common answer patterns and accordingly calculate the sentence or the passage with the best matching answer pattern for a given question pattern. There also exist some applications of QA in different domains. Abacha and Zweigenbaum (2015) propose an approach for the medical domain. The question classification step is completely pattern-based and uses pre-defined patterns and wh-words. The logical form of the question is created using a set of templates. After a query corresponding to the question is built, its several related forms are formed by removing some of the parts in order to increase the number of answers returned. The answers are ranked based on a justification metric.

2.2 Document summarization

The studies that review the field of text summarization analyze the problem based on several factors. Lloret and Palomar (2012) classify the research on summarization using the dimensions of summarization media (text, images, etc.), input form (single-document or multi-document), output content (extract, abstract, or headline), purpose of summaries (generic, query-focused, sentiment-based, etc.), and languages

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

involved (mono-lingual, multi-lingual, etc.). Based on the fact that the research on text summarization has focused mostly on generating extractive summaries rather than abstractive summaries, Nenkova and McKeown (2012) deal with extractive summaries only. They give a survey of topic representation, context-aware, graph-based, and machine learning approaches. The sentence selection process is explained in terms of greedy and global optimization algorithms, and the mechanisms to prevent duplicate sentences are given.

There are many recent studies on multi-document summarization and query-based summarization methods. In a work on query-oriented extractive multi-document summarization, Morita, Sakai and Okumura (2011) use a co-occurrence network of query words to formulate the summarization problem as a maximum coverage problem. The method is based on augmenting the query terms using the co-occurrence graph and then identifying the summary sentences. Zhong et al. (2015) use a deep learning framework for generating query-based summaries. The problem is solved in three stages formed of concept extraction, reconstruction validation, and summary generation. The experiments on benchmark document understanding conference (DUC) datasets showed around 0.38–0.43 Rouge-1 scores. The proposed method was compared with other representative summarization algorithms. Xiong and Ji (2016) use a hypergraph to identify the topic distribution in the sentences and learn the distribution between words and topics. The sentence selection process applies a random walk algorithm on the graph and the sentences are scored based on the diversity and centrality criteria. They obtained 0.42 Rouge-1 score on the same DUC dataset.

Wan (2009) designed a summarization method to be used in a QA system. It proposes a topic-based summarization framework. It differs from other works by taking into account the subtopics in a document in addition to the topic of the document. He argued that using relationships between sentences and subtopics increases the relevance of the summary. The method showed about 0.38–0.41 Rouge-1 success rates on DUC datasets. Another study that produces summaries as answers to questions uses reinforcement learning to determine the relevant features (Chali et al. 2015). A distinctive feature of the study is that the user can interact the system to select the best sentence among a number of candidate sentences at each iteration of the learning process. A rich set of features was used to determine the importance of sentences. The authors compared the method with SVM and k-means approaches.

In multi-document summarization, similarity measures are highly used in order to avoid choosing similar sentences for the summary (Mani 2001). According to Wang et al. (2012), similarity between sentences (sentence-sentence) and similarity between sentences and documents (sentence–document) can be used to find discriminative sentences in a set of documents. Discriminative sentence selection is an optimization problem on selecting an optimal subset of sentences. They employed a greedy approach assuming multivariate normal distribution and reported 0.53–0.57 Rouge-1 scores on blog and academic paper datasets. He et al. (2016) view a document as a signal formed of sentences. They argue that the signal is sparse in the sense that only a few sentences are important; these are the summary sentences. The sentences are grouped using a common pattern between the sentences. The sentence selection

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

process makes use of the groupings, thus avoiding selecting redundant sentences. They obtained 0.38 Rouge-1 score on the DUC 2006 dataset and compared the result with other participant systems. A similar work that uses a pattern-based model to reduce redundancy in sentence selection is given by Qiang et al. (2016). Alguliev, Alguliyev and Isazade (2013) formulate the summarization task in terms of an objective function based on coverage and diversity of the summaries. To increase the diversity, they model the problem using a differential evolution schema. They use modified versions of the cross-over and mutation operators. The Rouge-1 score was 0.39 on the DUC dataset.

Some studies approach the summarization problem from the perspective of event detection. Marujo et al. (2016) form multi-document summaries in terms of hierarchical combination of individual summaries. They developed two different methods named as single-layer summarization and waterfall summarization. An event detection method was proposed to identify irrelevant sentences (i.e., sentences that are not relevant to the main event) and also to determine similar events. The uninformative parts in the documents were eliminated based on the event representations. Glavas and Snajder (2014) convert documents into event graphs in which nodes correspond to event mentions and edges to temporal relations between events. They use a logistic regression classifier for determining the event words and a set of manually built rules to retrieve the arguments of events. The salient sentences are selected with respect to their relevance to the event mentions in the graph. Some other studies in multi-document summarization include those that focus on the sentence ordering problem (Bollegala, Okazaki and Ishizuka 2012) and that use more linguistic knowledge (Ferreira et al. 2014).

The main approach used in summarization for sentence selection is based on representing the sentences in terms of features and building a scoring schema on these features. Many different features such as location of the sentence, title, cue phrases, and occurrences of different words are used in these scoring schemas (Ferreira et al. 2013; Oliveira et al. 2016). Another frequently used and important method in summarization is based on lexical chains. Lexical chains semantically combine related terms across sentences and provide meaningful sequences throughout the text (Barzilay and Elhadad 1997; Silber and McCoy 2002; Li et al. 2007; Codina-Filba et al. 2017). Words represented by lexical chains are more informative compared to single words, thus they are quite helpful on revealing the concepts in the text and their inter-relations. Medelyan (2007) showed that a graph of disambiguated concepts shows a certain correlation with lexical chains.

3 System architecture

The overall pipeline of the HazırCevap system was designed in concordance with the DeepQA technology, introduced in IBM Watson (Ferrucci 2012). The primary principle is to have parallel pipelines with multiple submodules that produce different candidate results for each subproblem, which are then evaluated, scored, and prepared for the next module in the pipeline until the final answer is produced.

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

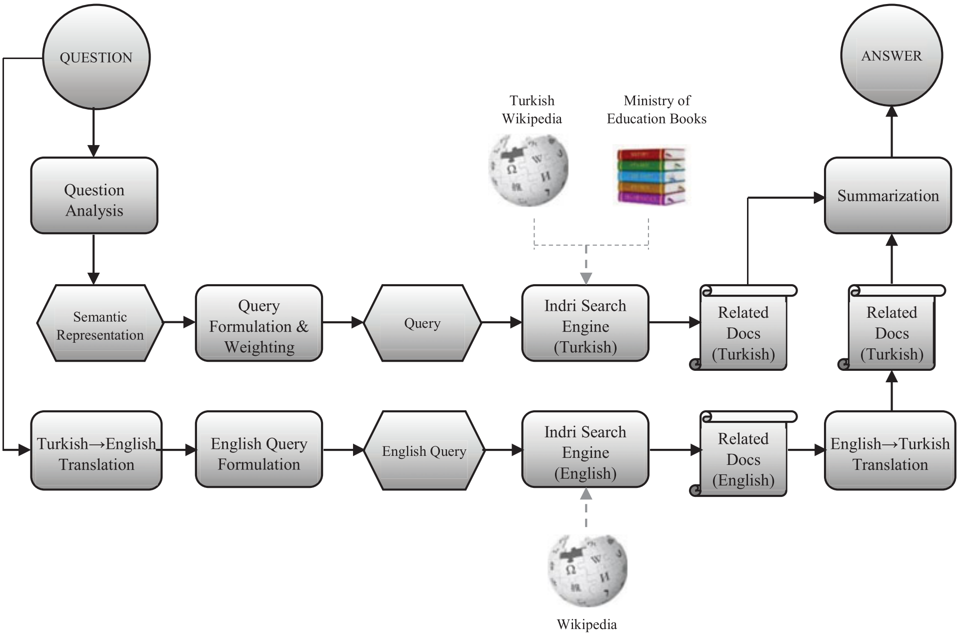

QUESTION
Question Analysis
Semantic Representation
Query Formulation & Weighting
Query
Turkish Wikipedia
Ministry of Education Books
Indri Search Engine (Turkish)
Related Docs (Turkish)
Summarization
ANSWER
Related Docs (Turkish)
Turkish→English Translation
English Query Formulation
English Query
Indri Search Engine (English)
Related Docs (English)
English→Turkish Translation
Wikipedia

Fig. 1. (Colour online) HazırCevap architecture.

The general system architecture is shown in Figure 1. As shown in the figure, HazırCevap has two main pipelines working in parallel to answer the given question. While the main pipeline is working with the Turkish resources, another pipeline works in parallel to reinforce the system using the English resources.

The question given to the system is firstly processed by the question analysis module, where the syntactically and semantically relevant informations are extracted. At this point, the system branches to the second pipeline to also work with the English resources, by translating the Turkish question into English. On both pipelines, the extracted data are used by the query formulation module to create the query for the IR module. The IR module collects the documents that are relevant to the given question from both Turkish and English resources. English documents are then translated into Turkish and appended to the original list of relevant documents. Then, the summarization module performs question-biased summarization on the relevant document collection to produce the final contextual answer.

4 Compilation of resources

In the first stage of system development, it was necessary to select a specific domain for an educational implementation of HazırCevap. The most important criterion for domain selection was that it had to be covered in the Turkish education curriculum. The second criterion was that the representation of information in a particular domain must be based more on verbal explanations rather than extensive formulas and/or diagrams and charts, and the third criterion was that resources that would be accessed through a search must be universal, more than local, and information must

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

be as objective as possible. By objective, we mean that the answers to a question should be similar in different resources.

After considering several options, Geography and Biology were selected as the two domains for the study, since these two met all of the three criteria for domain selection. Math and Chemistry, for example, which are also subjects found in the high school curriculum, failed the second criterion. As another example, History of the Turkish Republic failed the third criterion. The first working model of HazırCevap was created for Geography, and once the system worked properly in this domain, we added Biology as a second domain. The ultimate goal is to enhance the system so that it covers all of the subjects in the secondary and high school curriculum.

4.1 Pre-selection of resources

After the domains were selected, the next process was determining a set of reliable resources on which the system would operate. We downloaded ninth, tenth, eleventh and twelfth Grade Geography and Biology books as pdf files from Education Information Network (EBA)[6] of Institute of the Turkish Ministry of Education. These e-books were used as primary resources because they are entirely consistent with the curriculum, which is centrally prepared by the Ministry of Education. Based on the content covered in these e-books, we built a question database by writing 2,000 Geography questions and 2,000 Biology questions to be used in the resource selection process.

The questions were written by one researcher and two graduate students in the Educational Technology department. They were instructed to form questions based on the definitions of factoid and open-ended questions. In addition to the questions prepared by the research team, questions from the students in a high school were collected. A Google form was sent to the students by the teacher in the school and 300 student questions were collected. Table 1 shows example questions in the database.

In addition to this primary resource, a pilot study was carried out to determine a preliminary set of additional web resources to use in the HazırCevap system. 300 Geography and Biology questions were selected from our database of questions. We entered these 300 questions in the Google search engine[7] to locate resources relevant for each domain, without pre-evaluation criteria. As a result of searching the questions in Google, eighteen candidate resource websites were determined for Geography, and thirteen candidates were determined for the Biology domain.

4.2 Reliability and coverage metrics

The selection of websites used as web resources was based on a website evaluation checklist[8] developed by the University of California Berkeley Library. This checklist offers categories to check a website's reliability in several dimensions. We translated

[6] http://www.eba.gov.tr.

[7] http://www.google.com.

[8] http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/EvalForm_General.pdf

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Table 1. *Question examples*

Questions prepared by the research team
Geography factoid questions
Troposferin kalınlığına kadardır? (What is the thickness of the troposphere?)
Coğrafi keşifler hangi yüzyılda başlamıştır? (In which century have the geographical discoveries begun?)
Geography open-ended questions
Depremler ve tsunami arasında nasıl bir ilişki vardır? (What is the relationship between earthquakes and tsunami?)
Lejant (harita anahtarı) ne gibi kolaylıklar sağlar? (What kind of functions does the legend (map key) provide?)
Biology factoid questions
Biyosferin en büyük bölümünü kaplayan biyom çeşidi nedir? (What is the biome type that covers the largest part of the biosphere?)
Likenler hangi iki canlının ortak yaşamasıyla oluşur? (What are the two living beings that form lichens together?)
Biology open-ended questions
AIDS hastalığının bulaşma yöntemleri nelerdir? (What are the methods that cause infection of AIDS?)
Küresel iklim değişikliği nedir? (What is global climate change?)
Student questions
Geography factoid questions
Dicle nehriin başlangıç noktasıhangi şehirdedir? (Which city is the starting point of the Dicle river?)
En doğudaki ilimiz hangisidir? (Which is our easternmost city?)
Geography open-ended questions
Akarsuların aşınma olayınasıl gerçekleşir? (How does the erosion of rivers occur?)
Aurora ışıklarıneden sadece kutuplarda oluşur? (Why do the aurora lights only occur in the poles?)
Biology factoid questions
İnsanlarda kaç çift homolog kromozom vardır? (How many pairs of homologous chromosomes exist in humans?)
Aynıtür canlıların oluşturduğu topluluğa ne denir? (What is the community formed by the same species called?)
Biology open-ended questions
Canlılarda simbiyoz (birlikte yaşam) nedir? (What is symbiosis (common life) in living beings?)
Akraba evliliklerinin olasıriskleri nelerdir? (What are the possible risks of consanguineous marriages?)

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Table 2. *Reliability metrics*

Category	Subcategory	Explanation
Accuracy and authority	Domain type	Is the domain type (gov, org, com, etc.) reliable?
	Publisher	Is the publisher reliable?
	Personal	Is the information published on a blog or on a personal website?
	About	Does the "about" section give information about the author?
	Credentials	Are there enough credentials about the publisher?
Purpose and coverage	Links	Are the links on the page relevant to the topic? Are they all working?
	Well-organized	Is the website well-organized?
	Research findings	Does the page use scientific results as sources?
Design	Quality	How well is the page designed?

and adapted the list for our own use. The adapted version has three main categories and nine subcategories, as shown in Table 2.

The accuracy and authority category assesses the general reliability of the website by checking the type of the domain. For example, the "edu" domain name would be considered indication of a more reliable resource than a URL with a commercial (.com) domain name. Blogs or personal websites are not trusted resources since information there can be manipulated easily. The presence of an "about" section is an important indicator of a quality website, since it would provide information about the entity responsible for the content. The purpose and coverage category is about the content covered in a website and the organization of the information within the website. The links provided must be valid and complement of the topic of the website. References or evidence should be provided for any claims made on the website. Finally, the design category was kept on the evaluation rubric, since professional design and rational organization of information often denotes quality of a website.

We developed a scoring scheme to determine a reliability score, based on the rubric described above. A website is scored 1 point for each criterion it fully met, and a partial point (0.5) for a criterion partially met. If the website entirely failed the criterion, it received 0 points. The total score a website could collect was 9. Using the scoring scheme in Table 2, we ran the evaluation based on the 300 questions sampled from our question data bank. Among the eighteen websites for the Geography domain, we selected the top ten websites[9] and eliminated the rest.

[9] The websites with their contents and scores are: www.acikders.org.tr (course portal for basic and social sciences; 9.0), www.mgm.gov.tr (Turkish state meteorological service; 7.5), www.cografya.gen.tr (geography portal; 7.0), www.tr.wikipedia.org (Turkish wikipedia; 7.0), www.diyadinnet.com (news portal; 5.5), www.bilgiustam.com (portal about general domain; 5.0), www.nedirnedemek.com (Turkish dictionary; 5.0), www.turkcebilgi.com (a knowledge base in general domain; 4.5), www.konu-anlatimi.gen.tr (course portal for high school students; 3.5), www.bilgibirikimi.net (portal in general domain; 3.5).

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Table 3. *Web resources selected*

Geography web resources		Biology web resources	
www.eba.gov.tr	EBA website	www.eba.gov.tr	EBA website
www.cografya.gen.tr	Geography portal	www.tr.wikipedia.org	Turkish Wikipedia
www.tr.wikipedia.org	Turkish Wikipedia	www.biyolojidersnotlari.com	Biology course portal
www.diyadinnet.com	News portal	www.lisebiyoloji.com	Biology course portal
www.bilgiustam.com	Portal in general domain	www.webders.net	Course portal in general domain

In addition to the reliability check, the eighteen candidate websites were assessed in detail by a coverage scale based on the number of questions each can answer. We selected 200 reference questions from the question database. The sample was representative of the questions in terms of question classification. Using the Google search engine, each of the eighteen candidate websites were examined to determine how many questions it could answer. For each question answered, the website was given 1 point. The website would get 0 points if it failed to answer the question. In this way, the top ten websites[10] were selected related to the coverage metric in the Geography domain.

The compilation of resources was carried out in a similar fashion for the Biology domain. Then, the two measures, reliability and coverage, were superimposed for each domain and four websites with the most points were determined in each domain. The Education Information Network Institute's website (EBA) was not subjected to the evaluation rubric and scoring, since it is developed by the Ministry of Education, and contains the e-books used in schools that are the main information database of the study. Therefore, it was added as a fifth resource to the list by default. Table 3 shows the final list of resources for both Geography and Biology.

4.3 Identified resources and their compilation

The resource websites we identified for use in our database needed to be compiled as one set so that they could be added to the search engine index structure. Turkish content offered through Wikipedia was downloaded to be compiled offline. Wikipedia (Vikipedi in Turkish) presents all documents in separate pages. To use Vikipedi as a resource in HazırCevap, all Vikipedi pages were compiled according to document titles. The content between <*title*>…</*title*> tags was marked as the title of a document. The content of the document was obtained by investigating

[10] The websites with their contents and scores are: www.msxlabs.com (general forum and question answering site; 173), www.turkcebilgi.com (a knowledge base in general domain; 162), www.tr.wikipedia.org (Turkish wikipedia; 161), www.cografya.gen.tr (geography portal; 147), www.diyadinnet.com (news portal; 146), www.bilgiustam.com (portal in general domain; 136), www.forumdas.net (portal in general domain; 136), www.konu-anlatimi.gen.tr (course portal for high school students; 132), www.hakkinda-bilgi-nedir.com (portal in general domain; 130), www.bilgizenginleri.com (portal in general domain; 127).

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

the body tags <*body*>...</*body*> and the paragraph indicators <*p*>...</*p*>. This technique was applied to all Vikipedi documents and they were all compiled and added to the database.

All the e-textbooks on Biology and Geography offered on the EBA website were downloaded as pdf files. An open source program was used to extract plain text from the pdf documents, which was processed manually to determine its sections and subsections, indicated in the table of contents in each textbook. Each text block between two section/subsection titles were saved as separate documents. Since an OCR software was used to extract plain text from downloaded pdfs, all the text files were searched for possible spelling errors manually. Some extraneous material that pdf books contained such as graphics, tables, shapes, and quotations from newspapers were reorganized if they contributed to the topic.

5 Question analysis and representation

The first thing any QA system does to answer a question is to understand what the question is asking. Given the importance of this aspect of QA, question analysis in the scope of this work was comprehensively investigated on a separate study (Derici et al. 2015). The purpose of the question analysis module of HazırCevap is two-fold. First, it syntactically annotates the pieces of the question related to the answer and classifies the question into pre-defined question classes. Second (as an improvement to the original analysis), it produces a representation that characterizes the question and the answer by identifying the essential bits of information within the question. The results of both of these approaches prove to be useful for the other modules as well, therefore improving the overall effectiveness of the system.

Given a question, the question is first parsed using a Turkish dependency parser (Eryiğit, Nivre and Oflazer 2008). The accuracy of the parser is around eighty percent in terms of word-to-word attachment score. The analysis module then extracts the important pieces of information and produces a representation of the question. The pieces in the question that are of particular interest are the subject, the proper nouns, the focus, the mod(ifiers), and the class of the question. The *focus* of the question is defined as the set of terms in the question that indicate what type of entity is being asked for, and the *mod* is the collection of syntactic modifiers of the focus terms. The class is formed of a coarse class and a fine class (Derici et al. 2015). We identified seven coarse classes and fifty-six fine classes for Geography, and seven coarse classes and forty fine classes for Biology.

The focus identification is performed using two complementary approaches (Derici et al. 2014; Derici et al. 2015). A novel combination of a trained statistical hidden Markov model and a rule-based model is used to extract the focus terms. For the rule-based model, seven question types (e.g., what, how many) were identified and a rule (a pattern on the dependency parse tree) was formed manually for each that extracts the focus of the question. A pre-determined confidence score is given to each rule, which was calculated as the success of the rule on a training data set. The Hidden Markov model consists of two states (focus and non-focus) and performs a two-class classification on the sequence of words in the question using the Viterbi

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Table 4. *Scoring scheme of term types*

Term type	Score
Subject terms	1.5
Focus and mod terms	1.0
Proper nouns	2.0
Other terms	0.5

algorithm. In this way, each approach extracts candidate focus terms accompanied with their (normalized) scores. If a term is deemed as a focus term by both models, then it is accepted as a focus term. When the two models disagree on a term, the term is accepted as a focus term only if the score (determined by the accepting model) exceeds a threshold.

As in the case of focus identification, both rule-based and statistical classifiers were employed initially for the question classification problem. For the rule-based model, a set of phrases was determined for each class that are unique to that class. Given a question, the classifier assigns the question to the class with respect to the class words it includes. The statistical model employs a simple frequency-based approach. Given a question, it calculates, for each class, the tf-idf score of each word in the question. The word scores for each class are summed up and the question is assigned to the class with the highest score. In the current work, question classification is performed for coarse classes only. Classification for fine classes necessitates a sufficient amount of data for each class, which is left as future work. The experiments showed that the rule-based approach significantly outperforms the statistical approach. Thus, we use the rule-based question classification approach in the current work.

The mod terms are extracted from the dependency parse tree as the modifiers of the focus terms found. The subject of the question and the proper nouns are taken directly from the parse output. Finally, all the extracted information is combined to produce a representation of the given question, as shown below:

⟨Subject, Focus, Mod, QClass, PNouns⟩

The dependency parse tree of an example question is shown in Figure 2. Based on the information extracted from the tree and obtained from the focus and class identification models, the representation of the question is formed as follows:

⟨“ovası” (plain of), “ova” (plain), “en büyük” (the largest), ENTITY.PLAIN, “Türkiye” ⟩

After the analysis of the given question, the query formulation and weighting module produces the query to be used in the IR phase of the system. The query building module uses the representation of the given question for this purpose. Building the query begins with the separation of the subject terms, focus and mod terms, proper nouns, and other terms in the question. The other terms are then filtered by the question words (“what,” “which,” etc.) and the stopwords. Then, all the different groups of terms are scored according to the scoring scheme shown in Table 4 and a weighted query is formed accordingly.

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

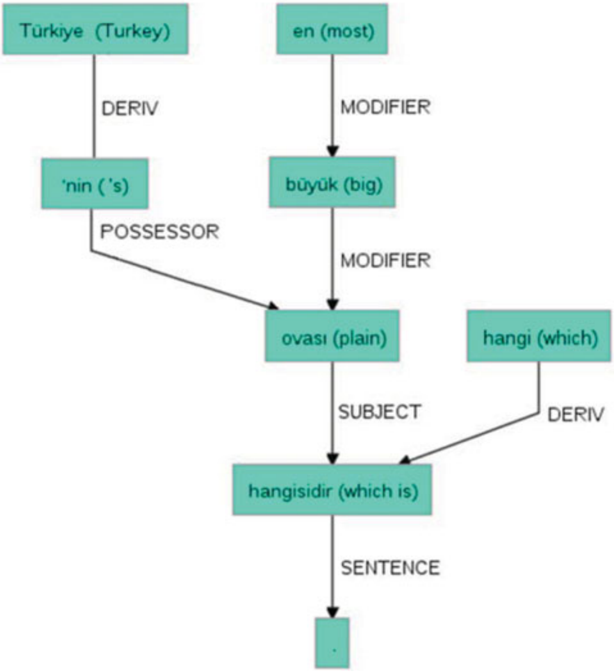


Fig. 2. (Colour online) Parse tree of “Türkiye’nin en büyük ovası hangisidir?” (“Which is the largest plain of Turkey?”).

These scores were determined by a series of manual experimentation over the range of 0.0–4.0. Each type has been fine-tuned both individually and relatively. For individual fine tuning, we isolated a particular type by setting all other types’ scores to 1.0, and set the best IR result of the type in interest. Relative tuning is performed by a series of manual experiments performed in a genetic algorithm fashion, i.e., eliminating the weak results and continue improving the tuning of the better scoring schemes. While there is the possibility of getting stuck at a local maximum, this approach proved to be effective for the test data set as well, which can also be seen in Section 7.

The scoring has proven to be sound also from the point of view of a search engine. For instance, the proper nouns generally have their own article in Wikipedia or their own section in the book resources. Therefore, having high relative scores for proper nouns improves the IR performance in returning more relevant documents to the question. On the other hand, focus and mod terms indicate a conceptual type of the answer, rather than being syntactically relevant terms (e.g., name of a plain). Therefore, too much increase in the score of these terms results in a decrease of the IR performance.

The query string generated by HazırCevap for the example question is shown below. In this query representation used by the Indri search engine, od_n and ud_n denote, respectively, an ordered set of terms and an unordered set of terms with a window size of n. That is, all the terms in the set should appear within a window of size n in the document in an ordered or unordered manner.

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

```
#weight (1.5 #ud1(ovası)
         1.0 #ud1(ova)
         1.0 #od2(en büyük)
         2.0 #ud1(Türkiye))
```

The works on question analysis in the literature usually focus on identifying the type or class of a question and employ a set of features specific to the class to retrieve candidate answers. IBM Watson follows a slightly different strategy and extracts entities such as the focus, modifiers, and lexical answer type of the question (Lally et al. 2012). In this work, we adapt a similar strategy. Rather than representing a question in terms of several types of features, we represent it using the focus, modifiers, subject, and proper nouns, which can be regarded as important features. The extraction of these entities is mostly based on the dependency parse of the question. In this respect, the methodology we use is similar to parse-based QA approaches to some extent.

6 Answer generation using document summaries

Previous research on QA has shown that users prefer getting some supplementary information in addition to the answer to the question (Mani 2001). Based on this, systems that provide additional information besides the answers to the questions aim at justifying the answer either by finding relevant passages in their corpora or by consulting to other resources such as the web. In HazırCevap, we follow a text summarization approach to respond to questions asked by students. The documents returned in response to a question are subjected to a multi-document summarization process and the summary is returned as the answer to the question. The summarization approach used here is based on an extracting technique, which is mainly concerned with choosing the top scoring sentences from the text in order to form the summary.

As in other summarization methods based on sentence extraction, we score the sentences in a document using a number of features (Ferreira et al. 2013; Oliveira et al. 2016). In this work, we use the three features explained below. The score of a sentence is the sum of the scores computed for each feature. In addition to using these features jointly, we also test the effects of different feature combinations via ablation studies (Section 7.2.2).

- *Term frequency*: Since the most frequent terms in a document give a general idea about the contents of the document, the term frequency metric is used in sentence scoring. In HazırCevap, we first identify the frequencies of all the terms in the document. We then take into account the terms whose frequencies are between a lower limit and an upper limit, and form a list of frequent terms for the document. The score of a sentence is incremented by 0.2 points for each term that exists in the frequent terms list.
- *Question words*: Since we aim at obtaining summaries that include the answer to a question, we give more importance to the words in the question compared to the other features. The score of a sentence is incremented by 1.0 points

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

for each word that also occurs in the question. In the preliminary tests, we observed that this feature is more effective than the other features, especially for factoid questions.

- *Lexical chains*: In the text summarization field, the lexical chain concept denotes a collection (chain) of important terms that occur in the documents (Li et al. 2007). Usually lexical ontologies such as WordNet are used to form the lexical chains corresponding to a document collection. Initially, we also considered WordNet for this purpose. However, we saw that Turkish WordNet is very limited for the Geography and Biology domains and we were unable to form meaningful chains. Therefore, in order to benefit from the method of lexical chains, we created two new ontologies that group related terms in these two domains. The ontology is formed of a set of groups. Each group consists of four entries; the first one is a general concept (title) and the others are subconcepts of this concept. For instance, the concept "kayaçlar" (rocks) and the subconcepts "magmatik" (magmatic), "tortul" (sedimentary) and "metamorfik" (metamorphic) form a group in the Geography ontology. The Geography and Biology ontologies contain sixty-five and eighty-one groups, respectively.

While using the lexical chain concept for sentence scoring, the term frequency feature was also taken into account. As a document is analyzed and the frequent terms are extracted, if such a term appears in a group in the ontology, we add all the four elements of that group to the lexical chain. Repeating this process for all the frequent terms in a document indicates that a lexical chain formed of frequent (important) and semantically related words is built. After the lexical chain for a document is formed, it is used in sentence scoring. For each term in a sentence that also appears in the lexical chain, the score of the sentence is incremented by 0.2 points. The logic behind this approach is that sentences that include domain-related terms may have important content.

For instance, assume that a document includes the word "magmatik" (magmatic) and this word is determined, by the term frequency metric, as a frequent word in the document. So, the words "kayaçlar" (rocks), "tortul" (sedimentary), and "metamorfik" (metamorphic) that are related to the word "magmatik" (magmatic) in the ontology are inserted to the lexical chain. After the lexical chain is formed, while scoring a sentence in the document, if the sentence contains one of these words in the lexical chain, 0.2 point is added to the sentence score for each such word.

The summarization algorithm is shown in Figure 3. Given a question and an input document, they are first subjected to a number of pre-processing steps. The stopwords are removed by using a stopwords list for Turkish. The root forms of the words are extracted using a Turkish morphological analyzer and disambiguator (Sak, Güngör and Saraçlar 2011). Then, the frequent terms list and the lexical chain are built. Following the initialization steps, sentences in the document are processed and scored, and the top scoring sentences are extracted to form the summary of the document. We use a pre-determined summary size depending on the length of the document. The size of the summary is ten percent of the original document if the

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

```
Algorithm SingleSummary
Input
  question: user question
  doc: document to be summarized
begin
  Preprocess question                   // stopwords removal and stemming
  Preprocess doc
  Divide doc into sentences sList
  for each distinct term t in doc       // build frequent terms list
    if (frequency of t is between lower and upper limits)
      Add t to freqTermsList
  endfor
  for each term t in freqTermsList      // build lexical chain
    if (t exists in a group in the Ontology)
      Add four elements of the group to LC
  endfor
  for each sentence s in sList          // process sentences
    for each term t in s
      if (t is in freqTermsList)
        Add 0.2 to score[s]
      if (t is in question)
        Add 1.0 to score[s]
      if (t is in LC)
        Add 0.2 to score[s]
    endfor
  endfor
  return top k sentences, where k is the summary size
end
```

```
Algorithm MultiSummary
Input
  sum_1,…,sum_n: single summaries
begin
  for each sentence $s_{1,j_1}$ in $sum_1$
   …
    for each sentence $s_{n,j_n}$ in $sum_n$
      Compute similarity of each pair of sentences in $\{s_{1,j_1},s_{2,j_2},\ldots,s_{n,j_n}\}$ using Eqn. (1)
      Identify the most similar pair as $s_{a,j_a}$ and $s_{b,j_b}$
      Mark $s_{b,j_b}$ in $sum_b$ as to be removed
    endfor
   …
  endfor
end
```

Fig. 3. Summarization algorithm. (a) Single document summarization. (b) Multi-document summarization.

document contains more than twenty sentences; twenty percent, if the number of sentences is between ten and twenty; and two sentences otherwise. The summary size for a single document is kept small since several summaries will be combined in the next phase. The summary sentences are listed in the order of their original positions in the document instead of the order of their scores. In this way, the cohesion of the document is maintained in the summary. Appendix A shows an example summary and comments on the use of the summarization features.

6.1 Multi-document summarization

After the individual summaries of the documents returned by the search engine are formed separately, they are combined and converted into a multi-document

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

summary by a novel approach (Figure 3). Suppose that the number of summary documents is n. Let $s_{i,j}$, $1 \leq i \leq n$, denote the jth sentence in the ith summary. The algorithm considers each group of n sentences $s_{1,j_1}, s_{2,j_2}, \ldots, s_{n,j_n}, \forall j_1 \ldots \forall j_n$. For a group of n sentences, it compares each pair of sentences and finds the similarity between each pair. That is, the similarities between the sentence pairs (s_{1,j_1}, s_{2,j_2}), (s_{1,j_1}, s_{3,j_3}),$\ldots$, $(s_{n-1,j_{n-1}}, s_{n,j_n})$ are computed. Then the pair with the maximum similarity score is identified and the second sentence in this pair is marked to be removed from the corresponding summary. The idea here is to eliminate one of the most similar sentences at each iteration. The algorithm continues until all the n-sentence groups are processed.

The similarity between two sentences is measured using a similarity metric similar to the cosine similarity:

$$\text{similarity} = \frac{k}{\sqrt{len_1 * len_2}} \tag{1}$$

where k is the number of common terms in the sentences, len_1 is the length of the first sentence, and len_2 is the length of the second sentence. In order not to damage the cohesion in the summaries, we keep the order of the sentences in each summary and output the summaries one by one in the final multi-document summary.

One point that should be noted here is that the time complexity of this process may seem high since all the sentences in all the documents are compared. However, we consider only a small number of documents (e.g., 3–5), and we work on the summaries of these documents already built rather than the original documents. In this respect, this sentence comparison process executes in a reasonable amount of time.

There are alternative methods used in the literature for avoiding redundancy in the final summary. One is the use of a metric such as maximal marginal relevance which, while adding a sentence to the summary, measures the similarity between that sentence and the already extracted summary sentences. A penalty factor is computed based on the similarity value. Another method is determining a threshold and, when comparing two sentences, eliminating one of them if their similarity exceeds this threshold. We also used this method with different threshold values in the initial phases of the work. However, we observed that determining such a threshold is a difficult issue and highly similar sentences may appear in the final summary depending on the threshold value. Thus, to prevent similarities in the output, we performed sentence removal based on relative similarities of sentence pairs to each other. In addition to measuring lexical similarity, there also exist methods based on semantic similarities of the words (e.g., Yang et al. 2016). A semantic similarity metric can be integrated into the multi-document summarization component explained in this section. We leave this extension as a future work.

7 Experiments and results

HazırCevap uses the Indri search engine for finding relevant documents of the given question. It is an open source search engine that utilizes language modeling and inference networks for the acquisition and sorting of the indexed documents (Metzler

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

and Croft 2004). Having a specially designed query language, Indri can index and directly query the documents formatted as pdf, HTML, and TREC. Additionally, it has a comprehensive API that supports several programming languages, making it considerably useful for systems like HazırCevap that utilizes different modules implemented in different programming languages. Furthermore, Indri has been proven to be useful for a lot of natural languages with different characteristics, even for morphologically complex languages such as Turkish.

Currently, HazırCevap has over 220,000 Turkish documents formatted as TREC text. TREC format is a standard XML-like format that is mostly used in related studies. We indexed the documents by both texts and titles. In the current pilot study, the document base includes all the Turkish Wikipedia articles along with the documents extracted from the books provided by the Turkish Ministry of Education.

In the English pipeline of the system, HazırCevap has 4.5M documents from English Wikipedia[11] indexed in the same manner as Turkish Wikipedia articles. The question given by the user is first translated to English using Google Translate API[12] before using the Indri search engine for finding relevant documents within these English documents. The top three relevant documents returned by Indri for the question are then translated into Turkish. We measured the success of the translations in both directions in a preliminary experiment. For the translation of questions from Turkish to English, the 400 questions (see Section 7.1) were translated and the outputs were evaluated using the five-point scale of Nagao, Tsujii and Nakamura (1988). For the translation of documents from English to Turkish, ten documents from each domain and question type (forty documents in total) were translated and evaluated with the same scale. The translation qualities were observed as eighty-two and eighty percent, respectively.

7.1 Question-document matching

Using the queries formed as explained in Section 5, we performed experiments to measure the success of retrieving documents that contain answers to the questions corresponding to the queries. For this purpose, hundred Geography factoid questions, hundred Geography open-ended questions, hundred Biology factoid questions, and hundred Biology open-ended questions were selected randomly from the question database. The selected questions were given to the search engine in the system and the top five documents returned by the search engine for each question were recorded.

In order to compare the success rate of HazırCevap with another system, the questions were also given to the Google search engine and similarly the top five documents were recorded. Two different experiments were conducted with Google. In the first one, while searching in Google, the phrase "site:tr.wikipedia.org" was used in the search bar to restrict the search to the Turkish Wikipedia site. For the second experiment, the EBA contents that were saved as separate documents (Section 4.3) were copied into folders in the project website (http://godel.cmpe.boun.edu.tr) and

[11] http://en.wikipedia.org
[12] https://cloud.google.com/translate/

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

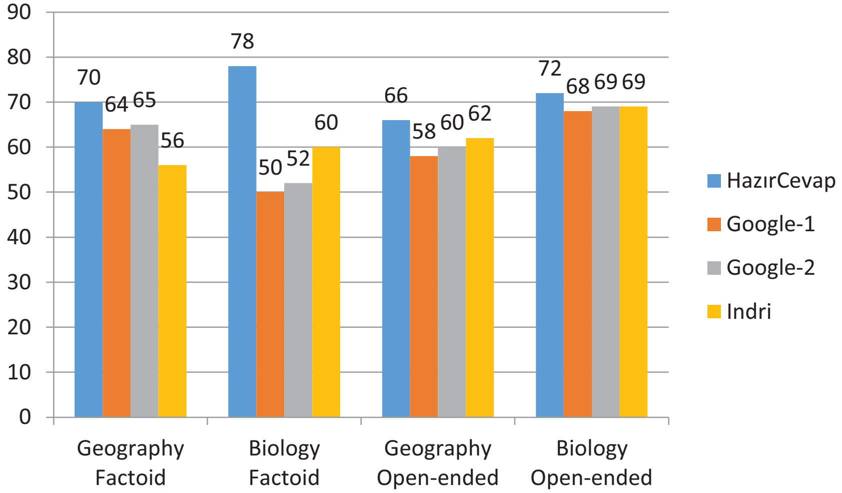


Fig. 4. (Colour online) Success rates of HazırCevap and search engines.

provided to be indexed by Google using indexing tools.[13] Then, during search, the phrase “site:tr.wikipedia.org OR site:godel.cmpe.boun.edu.tr” was used in the search bar to enable Google to use both Turkish Wikipedia and EBA resources. In this way, we aimed at searching on the same content in both the HazırCevap system and the Google search engine. In addition to Google, we also experimented with the Indri search engine, which is the search engine used in the proposed approach. The questions, without any processing, were given to Indri in exactly the same setting as HazırCevap. The purpose of this experiment was to observe the effect of question analysis on the success rate.

During the experiment, we investigated whether a document returned as a result of a question includes the answer to the question or not. This process is easier for factoid questions compared to open-ended questions. For a factoid question, by examining the content of a document, it was determined whether the answer can be found somewhere in the document or the answer can be deduced from the content. In other words, it was determined whether the person asking the question can arrive at the answer when he/she reads the document. A similar evaluation was done for open-ended questions. For these questions, however, it is ambiguous to some extent whether the document includes the answer or not. In this case, we considered whether the person asking the question can get an answer or explanation with sufficient details.

The results are shown in Figure 4. Google-1 denotes the results when only Turkish Wikipedia is used and Google-2 denotes the results using both Turkish Wikipedia and EBA resources (as in the HazırCevap system). The success rate of the HazırCevap system exceeds that of Google and Indri in both domains and in both question types. We measured the statistical significance of the results using the t-test. The performance of HazırCevap was observed to be significantly better

[13] https://www.google.com/webmasters/tools and http://www.google.com/addurl

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

($p < 0.05$) than the other three systems in the Biology factoid domain and the Indri system in the Geography factoid domain. The results lead us into two observations, in parallel to the results obtained in the pilot study in Section 7.3. The first one is that the success in the Biology domain is higher than the success in the Geography domain. When the question set is analyzed, we see that Biology questions include more domain-specific terms, while Geography questions include more general terms. This makes retrieving documents containing the specific terms given in the question easier in the Biology domain. The second observation is that the performance for factoid questions is higher than the performance for open-ended questions. This is an expected result. We search for an explicit and specific answer in factoid questions and this answer can be found in a related document. On the other hand, we search for more detailed answers and explanations in open-ended questions, and it is more difficult for a related document to contain a sufficient amount of information for the answer.

When we compare Google-1 and Google-2 results, we observe that including EBA resources does not contribute significantly to the success rate of the search engine. For both domains and question types, the increase in the number of answerable questions is just one or two. To understand why, we performed a test where hundred questions in each domain and question type were searched in Google using only the EBA resources (i.e., "site:godel.cmpe.boun.edu.tr"). Google was able to return at least one document for only six Geography factoid, ten Biology factoid, twenty-four Geography open-ended, and twenty-eight Biology open-ended questions. For the other questions (for example, for ninety-four Geography factoid questions), no documents were returned. When we analyzed the reason of this case, we observed that when a question in natural language is fed to the search engine without any processing (except some processing interior to the search engine), it is highly difficult for the search engine to find documents that include all the contents in the question. As the number of terms in the question increases, the chance of returning answers decreases. This is an expected result. On the other hand, as done in the HazırCevap system, analyzing a question and giving some parts of the question to the search engine after filtering the rest increases the chance of finding relevant documents.

In addition to the comparison with other search engines, we also examined the performance of the system on the top k documents ($1 \leq k \leq 5$). Figure 5 shows the results. As can be seen in the figure, for questions that can be answered, the answer appears mostly in the top first or second documents. The remaining documents do not contribute much to the answer. For instance, for Biology open-ended questions, fifty-eight of the questions were answered by the top document returned by the search engine. Including the second documents increases this number to just sixty-six. This means that for only eight documents, the answer does not appear in the first document but appears in the second document. The contribution to the answer was decreased more for the documents after the top two documents.

7.2 Document summarization

We conducted a document summarization experiment to check whether or not the document summaries returned by the HazırCevap system contained the answer, to

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

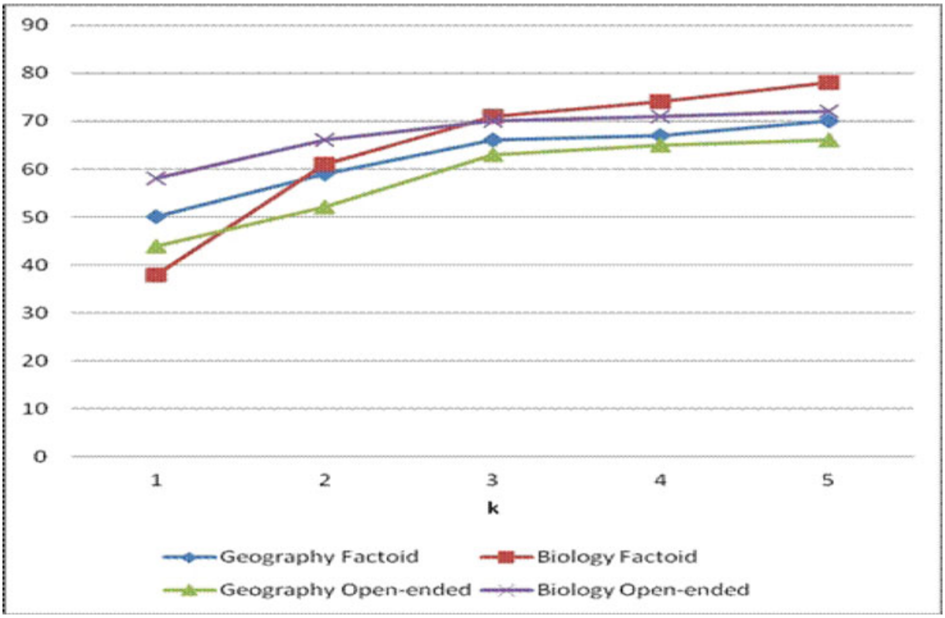


Fig. 5. (Colour online) Success rates for top k documents.

determine the extent to which the summaries were relevant for the question entered, and to measure how much the system summaries correlate with human summaries. We used the same 400 questions used in the previous experiment. The subsections below explain first the manual evaluation of the summaries and then the automatic evaluation results.

7.2.1 Manual evaluation

The user evaluation was conducted in two phases. In the first phase, two annotators decided if the answers to the questions were found in the document summaries by a yes/no scoring scheme. (This is similar to the experiment in Section 7.1; the difference here is that document summaries are evaluated rather than the documents themselves.) The results are shown in Table 5. The scores of the annotators and also the average scores indicate that over fifty percent of the answers to the 400 questions were found in the summaries returned by HazırCevap. We also measured the agreement between the two raters using the Cohen's kappa coefficient. The agreement was calculated as 0.73 in this phase. Landis and Koch (1977) stated that if Cohen's kappa coefficient is within the range of 0.61–0.80, the strength of the agreement is substantial. The annotators' agreement can be considered substantial (0.73) in our experiment.

In the second phase of the experiment, points from 1 to 5 were given to assess the extent to which the summaries were relevant to the contents of the questions. By relevance to a question, we mean the amount of material in the summary that is related to the information need stated in the question. The percentage scores of relevance of the summaries for the questions are shown in Figure 6. For factoid

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Table 5. *Success rates for the summaries*

	1st Annotator		2nd Annotator		Average	
%	Yes	No	Yes	No	Yes	No
Geography factoid	55	45	58	42	56.50	43.50
Geography open-ended	52	48	44	56	48.00	52.00
Biology factoid	54	46	49	51	51.50	48.50
Biology open-ended	64	36	53	47	58.50	41.50
Average	56.25	43.75	51.00	49.00	53.63	46.37

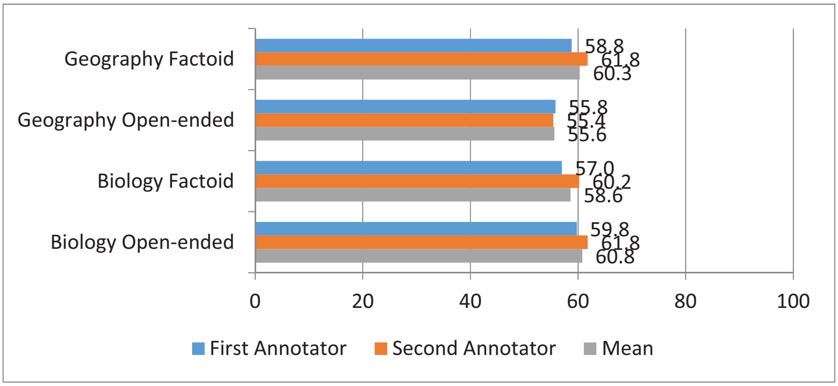


Fig. 6. (Colour online) Percentages of relevance to the questions' contents.

Geography questions, the mean percentage was approximately sixty percent, and for open-ended questions it was fifty-six percent. Similarly, for factoid Biology questions the relevance was fifty-nine percent, while open-ended Biology questions scored the highest mean percentage, sixty-one percent According to these analyses, we can conclude that the summaries of the HazırCevap system were related to questions' contents in the ratio of approximately fifty-nine percent.

In order to test the effect of question analysis on the overall response of HazırCevap, we performed an additional experiment. The questions were given to the Indri search engine without any question analysis and the retrieved documents were summarized. The two phases of the user evaluation were repeated with one annotator. The results are shown in Table 6. When compared with the corresponding figures in Table 5 and Figure 6, we observe that question analysis is a significant component in the HazırCevap system.

7.2.2 Automatic evaluation

In addition to manual evaluation, we measured the success of the summaries prepared by the system by comparing them to manual summaries using the Rouge metric. The summaries were evaluated with respect to the Rouge-n ($n = 1, 2$) and Rouge-SU4 metrics (Lin 2004). The experiment was performed using ten

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Table 6. *Success rates and relevance without question analysis*

	Success rate		Relevance
%	Yes	No	to question
Geography factoid	49	51	50.2
Geography open-ended	44	56	45.2
Biology factoid	44	56	44.0
Biology open-ended	52	48	46.2
Average	47.25	52.75	

Table 7. *Evaluation results of summaries for Rouge-n (n=1,2)*

	Rouge-1			Rouge-2		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Geography factoid	0.23	0.59	0.33	0.17	0.44	0.25
Geography open-ended	0.28	0.51	0.36	0.18	0.34	0.24
Biology factoid	0.33	0.45	0.38	0.21	0.28	0.24
Biology open-ended	0.40	0.46	0.43	0.27	0.30	0.28
Average	0.31	0.50	0.38	0.21	0.34	0.26

questions selected randomly from the set of hundred questions for each domain and question type. For each question, the reference summary is formed by taking the first five documents returned by the HazırCevap system, manually summarizing them separately, and combining these summaries one after another. The reference summaries were prepared by one annotator outside the project team. For each question, he is given the question and the five documents. He is requested to extract (if any) from each document the answer to the question and the parts that are relevant to the contents of the question. Since the lengths of the system summaries are not fixed, he is allowed to extract parts of any length as he would like to see as a summary of the question.

Rouge is basically a recall-oriented metric. In this research, we used the forms that use both recall and precision (Lin 2004; Ganesan, Zhai and Han 2010). The results are shown in Tables 7 and 8. Each entry in the tables is the average of the ten questions for that domain and question type. The first point that draws attention about the results is the difference between the recall and precision values. The reason is that the reference summaries are shorter than the system summaries. We observed that while preparing manual summaries people filter the documents and select only sentences that are relevant with the question. On the other hand, the summaries formed by the system might include sentences that are not relevant but that pass the summarization criteria (e.g., including question terms).

When Rouge-1 and Rouge-2 scores are compared, we see a decrease in Rouge-2 scores as expected. This indicates that, in Rouge-1, some of the terms common to

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Table 8. *Evaluation results of summaries for Rouge-SU4*

	Rouge-SU4		
	Precision	Recall	F-measure
Geography factoid	0.17	0.43	0.24
Geography open-ended	0.19	0.35	0.25
Biology factoid	0.23	0.31	0.27
Biology open-ended	0.28	0.32	0.30
Average	0.22	0.35	0.27

the reference and system summaries (unigrams) do not actually occur in the same sentence in the two summaries. As bigrams are also taken into account in Rouge-2, it is understood that some common terms that contribute to the Rouge-1 score do not occur as common bigrams in the two summaries.

We can also compare the results with respect to the domain and the question types. In all the measures in this section, Biology results are higher than Geography results. This is probably due to the fact that Biology questions include more domain-specific terms compared to the Geography questions. Thus, the system summaries in the Biology domain contain more sentences relevant to the content of the question. As will be discussed in Section 7.3, this finding is also supported by the pilot study.

In order to observe the effect of each of the three features used in the summarization process, we also performed ablation studies by switching on or off each feature type. Table 9 shows the results. We see that the question words feature is the most important feature type and the other features have less effect on the success rate. When all the features are used, the success is improved in one domain (Biology open-ended) compared to using the question words feature, and the performances are similar in the other domains.

As a result, we conclude that 0.33–0.43 Rouge-1 F-measure (0.38 on average), 0.24–0.28 Rouge-2 F-measure (0.26 on average) and 0.24–0.30 Rouge-SU4 F-measure (0.27 on average) success rates were obtained in different domains and question types.[14]

When we analyze the errors caused by the system, we see that the most critical errors occur during question analysis. The errors originate from two components: the dependency parser and the module that extracts the important parts in the question. When the parser errors affect parts in the dependency tree that are used in extracting information, incorrect data are extracted. Even when the parser output is correct, the analysis module may yield wrong pieces of information. The following questions exemplify these two cases:

[14] It is usually difficult to compare the Rouge figures in text summarization studies. The evaluation depends on several factors such as the domain, the language, the number of reference summaries per document, and the summary length. Some state-of-the-art results were mentioned in Section 2.2. We can also cite a recent study, in which a comprehensive analysis was presented that compares eighteen extractive summarization methods for both single-document and multi-document summarization (Oliveira et al. 2016). The Rouge-1 scores obtained were in the ranges of 0.22–0.53 for single-document summarization and 0.20–0.37 for multi-document summarization.

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Table 9. *Ablation study evaluation results for Rouge-1 (TF: term frequency, QW: question words, LF: lexical chains)*

	TF			QW			LC			TF-QW			TF-LC			QW-LC		
	Pr	Rec	F	Pr	Rec	F	Pr	Rec	F	Pr	Rec	F	Pr	Rec	F	Pr	Rec	F
Geography factoid	0.14	0.66	0.23	0.23	0.62	0.33	0.39	0.24	0.30	0.23	0.59	0.34	0.14	0.66	0.23	0.23	0.62	0.33
Geography open-ended	0.21	0.60	0.31	0.27	0.56	0.36	0.34	0.15	0.21	0.28	0.51	0.36	0.21	0.60	0.31	0.27	0.56	0.36
Biology factoid	0.26	0.54	0.35	0.30	0.56	0.39	0.44	0.14	0.21	0.33	0.45	0.38	0.26	0.54	0.35	0.30	0.56	0.39
Biology open-ended	0.29	0.51	0.37	0.48	0.31	0.38	0.53	0.13	0.21	0.40	0.46	0.42	0.29	0.51	0.37	0.48	0.31	0.38
Average	0.22	0.58	0.32	0.32	0.52	0.39	0.43	0.17	0.24	0.31	0.50	0.38	0.22	0.58	0.32	0.32	0.52	0.39

“Avustralya mercan resiflerinin uzunluğu ne kadardır?” (“What is the length of the Australian coral reefs?”)

“Belirli bir bitkiden aynı kalıtsal yapıda ikinci bir bitki elde etmek için nasıl bir yol izlenebilir?” (“What type of a method must be followed to obtain a new plant from a specific plant with the same hereditary structure?”)

In the first one, the subject is identified incorrectly as “Avustralya” (“Australian”) by the parser. The reason of the error is probably the use of the word “Avustralya” in both non-possessive sense (“Australia”) and possessive sense (“Australian”) in Turkish. In the second one, the analysis outputs the word “bitki” (“plant”) instead of the word “yol” (“method”) as the focus. The terms were identified incorrectly in such cases causing improper assignment of weights to the Indri query and thus decreasing the retrieval performance.

7.3 Pilot study

We conducted a pilot testing of the HazırCevap system with users in a high school in Istanbul. As stated previously, the system was implemented for use in the domains of Biology and Geography, both of which are mandatory subjects in tenth grade education in Turkey. A total of thirty-three 10th graders tested the system in two groups: sixteen students were in the Geography group, and seventeen in the Biology group. We asked the students to formulate one open-ended and two factoid questions in the domain to which they were assigned.

The system was made available to the students over the web, and the testing was conducted in the school’s computer lab. At the beginning of each testing session, one of the researchers briefly demonstrated how HazırCevap works and illustrated factoid and open-ended types of questions. Then, the students formulated and entered questions in HazırCevap. Table 10 shows example questions written by the students for each domain. Since the students were shown factoid and open-ended question examples at the beginning of the session, the questions they wrote were well-formed and there were not significant differences from those in the question database (Table 1). The major difference was the length of the questions; students mostly preferred writing shorter questions.

The students evaluated the answers returned by the system by filling out a short evaluation form for each answer. The form had three questions for each system response and two general usability questions. For each question’s answer, the students were first asked whether or not the answer they received was relevant and then to judge on a Likert scale (1–5) the degree of relevance of the information provided in the Turkish summary as well as the translated summary. For the general evaluation of the system, the students were also asked to comment on system usability: the ways in which HazırCevap would be useful to them, and what was needed to make the system more beneficial.

As shown in Table 11, the students indicated that the system returned a relevant answer in Biology for eighty-eight percent of the factoid questions and eighty-two percent of the open-ended questions. In contrast, only seventy percent of the factoids

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Table 10. *Question examples in the pilot study*

Geography factoid questions
Akifer nedir? (What is an aquifer?)
Dünyanın en büyük yüzölçümü hangi ülkededir? (Which country has the largest area in the world?)
Geography open-ended questions
Toprak oluşumunu etkileyen faktörler nelerdir? (What are the factors that affect pedogenesis?)
Nüfus yoğunluğu nedir? (What is population density?)
Biology factoid questions
Bitkilerin tozlaşmasınısağlayan besin nedir? (What is the nutrient that helps pollination in plants?)
Kaç tip üreme tipi vardır? (How many reproduction types are there?)
Biology open-ended questions
Mutasyon nasıl olur? (How does mutation occur?)
Dünyada bulunan ekosistem çeşitleri nelerdir? (What are the ecosystem types in the world?)

Table 11. *Student evaluation of the HazırCevap system*

	Biology		Geography	
%	Factoid	Open-ended	Factoid	Open-ended
Did the system return a relevant answer?	88	82	70	56
Was the Turkish summary relevant to your question?	86	79	70	58
Was the translated summary relevant to your question?	74	79	60	43

and a mere fifty-six percent of the open-ended questions were answered in the domain of Geography. The judgments of relevance of the summaries also differed based on the domain. The students responded that the summary from the Turkish resources was relevant in eighty-six percent, and the translated summary from the English resources was relevant in seventy-four percent of the factoid questions in Biology. This ratio was seventy-nine percent for both types of summaries for open-ended questions in Biology. The students indicated that for the factoids in Geography, the relevance of the Turkish summary was seventy percent, while that of the translated summary was sixty percent. As for the open-ended questions in Geography, the ratio dropped to fifty-eight percent for the Turkish summary and forty-three for the translated summary.

When we examined the surface features of the students' questions, we detected a stark difference in the composition of both factoid and open-ended questions based on domain. While twenty-five out of thirty-four factoid questions in Biology

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

contained a domain-specific term, only ten of the thirty-two questions contained a term unique to Geography. Similarly, fourteen out of the seventeen open-ended Biology questions had a unique jargon, while only three of the sixteen Geography questions made use of specific terminology. We suspect that the large difference in the students' judgments for each domain might be a result of the composition of the questions with or without domain jargon. We also analyzed the differences in the success rates between Turkish and English summaries. The main reason of the decrease in performance in English summaries seems to originate from the quality of translations. As mentioned in Section 7, there is a degradation in the accuracy and fluency of the questions and documents during the translations in both directions. This introduces ungrammatical sentences as well as irrelevant information in the output summaries.

As for the findings regarding usability, three types of possible usage were suggested by the students. A majority of the students (seventy-seven percent) responded that they would use HazırCevap in order to study for an exam or when doing their homework, while more than half (fifty-three percent) indicated that the system would be useful for finding general information. A smaller percentage of the participants (ten percent) said such a system would be most useful for speedy access to information on demand. The students' suggestions for increasing usability fell into three categories: Content, speed, and interaction design. Generally, the students requested that the system be improved so that it can respond to all types of questions, provide the answer faster, and encompass all of the subjects in the high school curriculum. Finally, some of the students suggested that the interface can be designed in a more professional way, as in search engines such as Google.

8 Conclusion

In this paper, we developed a Turkish QA system named as HazırCevap. Our main motivation was to build a framework that can provide students reliable and accurate answers related to their questions on their subject studies such as Geography and Biology.

In the QA approach proposed in this work, we aimed for accuracy, high coverage, and sufficiency. Related to accuracy, we conducted reliability checks on the web resources using rubrics from credible institutes. The whole study, from methodological design up to evaluation, was guided by a team of researchers from Educational Technology department in order not to miss the educational point of view of the study. For high coverage, we integrated web resources such as Wikipedia, educational web sites with digital copies of textbooks, and multi-lingual resources. Finally, for the sufficiency requirement, in addition to answering the students' questions, we designed the system to also supply the student's important details about the question topic and related documents for further investigation of the subject.

We performed several experiments to measure the effectiveness of the methods developed. After the questions were converted into question representation, about 65%–80% of the documents returned by the search engine contained the answers to

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

the questions. To compare the results with another system, the questions were also given to the Google search engine without any preprocessing. The success rate of Google was found to be about ten percent lower than HazırCevap. We also tested the system as a whole by considering the summaries formed as answers to questions. About fifty percent of the summaries included answers to the questions and about 55%–60% of the summaries were relevant to the topics of the questions.

Although HazırCevap was designed for Turkish students, it is not a language dependent system. It can easily be used in other languages. The system can be adapted to any language and domain having the required tools and resources without any further training. Our experiments showed that HazırCevap performs better on subject domains with a higher percentage of special terms than domains that include more general terms. We have prepared the first question dataset for Turkish having 4,000 factoid and open-ended questions in the domains of Geography and Biology. The dataset consists of the questions, their answers, focus, modifier, and class information. The dataset is publicly available at the project website.[15]

The research in this paper has a number of theoretical and practical implications. From the theoretical perspective, it offers a method to compile a set of reliable resources, provides a question representation based on structural analysis, and proposes a QA schema that is formed of multi-document summaries of both native language and foreign language documents. The practical implication is the adaptation of the framework in an educational setting. The pilot study of the developed system in a high school environment showed that the approach is substantially useful for obtaining answers and getting relevant information in the summaries.

As future work, we plan to extend HazırCevap to work on other subjects like History, Chemistry, etc. This will require developing new modules for parsing and processing graphical data and equations. Another future work is related to the question analysis component. In the current work, by analyzing the patterns in the question dataset, we determined a number of possible question patterns. Some of the questions cannot be handled by the currently available patterns. We plan to increase the question types that can be parsed and identify the patterns for these types. As another extension, the ontologies built within the scope of this work should be extended in order to increase the performance of the summarization component. Finally, the sentence similarity metric used in the summarization module can be improved by incorporating semantic similarity of the words.

References

Abacha, A.B., and Zweigenbaum, P. 2015. MEANS: a medical question-answering system combining NLP techniques and semantic web technologies. *Information Processing and Management* **51**: 570–94.

Alguliev, R.M., Aliguliyev, R.M., and Isazade, N.R. 2013. Multiple documents summarization based on evolutionary optimization algorithm. *Expert Systems with Applications* **40**: 1675–89.

[15] http://godel.cmpe.boun.edu.tr/Public.

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Barzilay, R., and Elhadad, M. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 10–7.

Bernhard, D., and Gurevych, I. 2009. Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *Proceedings of ACL-IJCNLP*, pp. 728–36.

Bollegala, D., Okazaki, N., and Ishizuka, M. 2012. A preference learning approach to sentence ordering for multi-document summarization. *Information Sciences* **217**: 78–95.

Bordes, A., Chopra, S., and Weston, J. 2014. Question answering with subgraph embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 615–20.

Bordes, A., Weston, J., and Usunier, N. 2014. Open question answering with weakly supervised embedding models. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, Springer-Verlag, pp. 165–80.

Bouziane, A., Bouchina, D., Doumi, N., and Malki, M. 2015. Question answering systems: survey and trends. *Procedia Computer Science* **73**: 366–75.

Brill, E., Dumais, S., and Banko, M. 2002. An analysis of the AskMSR Question-Answering system. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 257–64.

Chali, Y., Hasan, S.A., and Mojahid, M. 2015. A reinforcement learning formulation to the complex question answering problem. *Information Processing and Management* **51**: 252–72.

Chen, Y., Zhou, M., and Wang, S. 2006. Reranking answers for definitional QA using language modeling. In *Proceedings of ACL/COLING*, pp. 1081–8.

Chu-Carroll, J., Fan, J., Boguraev, B.K., Carmel, D., Sheinwald, D., and Welty, C. 2012a. Finding needles in the haystack: search and candidate generation. *IBM Journal of Research and Development* **56**(3): 300–11.

Chu-Carroll, J., Fan, J., Schlaefer, N., and Zadrozny, W. 2012b. Textual resource acquisition and engineering. *IBM Journal of Research and Development* **56**(3/4): 4.1-4.11.

Codina-Filba, J., Bouayad-Agha, N., Burga, A., Casamayor, G., Mille, S., Müller, A., Saggion, H., and Wanner, L. 2017. Using genre-specific features for patent summaries. *Information Processing and Management* **53**(1): 151–74.

Derici, C., Çelik, K., Kutbay, E., Aydın, Y., Güngör, T., Özgür, A., and Kartal, G. 2015. Question analysis for a closed domain question answering system. In A. Gelbukh (ed.), Proceedings of Computational Linguistics and Intelligent Text Processing (CicLing), pp. 468–82. Springer, Cairo.

Derici, C., Çelik, K., Özgür, A., Güngör, T., Kutbay, E., Aydın, Y., and Kartal, G. 2014. Türkçe soru cevaplama sistemlerinde kural tabanlıodak çıkarımı(Rule-based focus extraction in Turkish question answering systems). In *Proceedings of Signal Processing and Communications Applications Conference (SIU)*, pp. 1604–7.

Diefenbach, D., Lopez, V., Singh, K., and Maret, P. 2017. Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information Systems*, pp. 1–41, Berlin, Germany: Springer.

Dong, L., Wei, F., Zhou, M., and Xu, K. 2015. Question answering over Freebase with multi-column convolutional neural networks. In *Proceedings of International Joint Conference on Natural Language Processing (IJNLP)*, pp. 260–9.

Er, N.P., and Çiçekli, I. 2013. A factoid question answering system using answer pattern matching. In *Proceedings of International Joint Conference on Natural Language Processing (IJNLP)*, pp. 854–8.

Eryiğit, G., Nivre, J., and Oflazer, K. 2008. Dependency parsing of Turkish. *Computational Linguistics* **34**(3): 357–89.

Fan, J., Kalyanpur, A., Gondek, D.C., and Ferrucci, D.A. 2012. Automatic knowledge extraction from documents. *IBM Journal of Research and Development* **56**(3/4): 5:1–5:10.

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Feng, M., Xiang, B., Glass, M.R., Wang, L., and Zhou, B. 2015. Applying deep learning to answer selection: a study and an open task. In *Proceedings of Automatic Speech Recognition and Understanding (ASRU)*, pp. 813–20.

Ferreira, R., Cabral, L.de S., Freitas, F., Lins, R.D., Silva, G.de F., Simske, S.J., and Favaro, L. 2014. A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications* **41**: 5780–7.

Ferreira, R., Cabral, L. de S., Lins, R.F., Silva, G.P., Freitas, F., Cavalcanti, G.D.C., Lima, R., Simske, S.J., and Favaro, L. 2013. Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications* **40**: 5755–64.

Ferrucci, D.A. 2012. Introduction to “this is Watson”. *IBM Journal of Research and Development* **56**(3): 235–49.

Figueroa, F., and Neumann, G. 2016. Context-aware semantic classification of search queries for browsing community question–answering archives. *Knowledge-Based Systems* **96**: 1–13.

Ganesan, K., Zhai, C.X., and Han, J. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of International Conference on Computational Linguistics (COLING)*, pp. 340–8.

Glavas, G., and Snajder, J. 2014. Event graphs for information retrieval and multi-document summarization. *Expert Systems with Applications* **41**: 6904–16.

Gondek, D.C., Lally, A., Kalyanpur, A., Murdock, J.W., Duboue, P.A., Zhang, L., Pan, Y., Qiu, Z.M., and Welty, C. 2012. A framework for merging and ranking of answers in DeepQA. *IBM Journal of Research and Development* **56**(3): 399–410.

Habibi, M., Mahdabi, P., and Popescu-Belis, A. 2016. Question answering in conversations: query refinement using contextual and semantic information. *Data & Knowledge Engineering* **106**: 38–51.

He, R., Tang, J., Gong, P., Hu, Q., and Wang, B. 2016. Multi-document summarization via group sparse learning. *Information Sciences* **349–50**: 12–24.

Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J., and Ngomo, A.-C.N. 2016. Survey on challenges of question answering in the semantic web. *Semantic Web* **8**(6): 1–26.

Iyyer, M., Boyd-Graber, J., Claudino, L., Socher, R., and Daume, H. 2014. A neural network for factoid question answering over paragraphs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 633–44.

İlhan, S., Duru, N., Karagöz, Ş., and Sağır, M. 2008. Metin madenciliği ile soru cevaplama sistemi (A question answering system based on text mining). In *Proceedings of Elektrik-Elektronik ve Biyomedikal Mühendisliği Konferansı(ELECO) (Conference on Electrical-Electronics and Biomedical Engineering)*, pp. 356–9.

Katz, B. 1997. Annotating the world wide web using natural language. In *Proceedings of the Conference on Computer Assisted Information Searching on the Internet (RIAO)*, pp. 136–55.

Khodadi, I., and Abadeh, M.S. 2016. Genetic programming-based feature learning for question answering. *Information Processing and Management* **52**: 340–57.

Kolomiyets, O. and Moens, M.F. 2011. A survey on question answering technology from an information retrieval perspective. *Information Sciences* **181**: 5412–34.

Lally, A., Prager, J.M., McCord, M.C., Boguraev, B.K., Patwardhan, S., Fan, J., Fodor, P., and Chu-Caroll, J. 2012. Question analysis: how Watson reads a clue. *IBM Journal of Research and Development* **56**(3/4), 2:1–2:14.

Landis, J.R., and Koch, G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics* **33**(1): 159–74.

Li, J., Sun, L., Kit, C., and Webster, J. 2007. A query-focused multi-document summarizer based on lexical chains. In *Proceedings of the Document Understanding Conference (DUC).*

Lin, C.-Y. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization Branches Out (WAS)*, pp. 74–81.

Lloret, E. and Palomar, M. 2012. Text summarisation in progress: a literature review. *Artificial Intelligence Review* **37**(1): 1–41.

Mani, I. 2001. *Automatic Summarization.* Amsterdam: John Benjamins Pub.

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Marujo, L., Ling, W., Ribeiro, R., Gershman, A., Carbonell, J., de Matos, D.M., and Neto, J.P. 2016. Exploring events and distributed representations of text in multi-document summarization. *Knowledge-Based Systems* **94**: 33–42.

McCord, M.C., Murdock, J.W., and Boguraev, B.K. 2012. Deep parsing in Watson. *IBM Journal of Research and Development* **56**(3/4), 3-1:3-15.

Medelyan, O. 2007. Computing lexical chains with graph clustering. In *Proceedings of the Annual Meeting of the ACL: Student Research Workshop*, pp. 85–90.

Metzler, D., and Croft, W.B. 2004. Combining the language model and inference network approaches to retrieval. *Information Processing and Management* **40**(5): 735–50.

Mishra, A., and Jain, S.K. 2016. A survey on question answering systems with classification. *Journal of King Saud University* **28**: 345–61.

Molino, P., Lops, P., Semeraro, G., Gemmis, M., and Basile, P. 2015. Playing with knowledge: a virtual player for "who wants to be a millionaire?" that leverages question answering techniques. *Artificial Intelligence* **222**: 157–81.

Momtazi, S. and Klakow, D. 2015. Bridging the vocabulary gap between questions and answer sentences. *Information Processing and Management* **51**: 595–615.

Morita, H., Sakai, T., and Okumura, M. 2011. Query snowball: a co-occurrence-based approach to multi-document summarization for question answering. In *Proceedings of the Annual Meeting of the ACL*, pp. 223–9.

Murdock, J.W., Fan, J., Lally, A., Shima, H., and Boguraev, B.K. 2012a. Textual evidence gathering and analysis. *IBM Journal of Research and Development* **56**(3): 325–38.

Murdock, J.W., Kalyanpur, A., Welty, C., Fan, J., Ferrucci, D.A., Gondek, D.C., Zhang, L., and Kanayama, H. 2012b. Typing candidate answers using type coercion. *IBM Journal of Research and Development* **56**(3): 312–24.

Nagao, M., Tsujii, J., and Nakamura, J. 1988. The Japanese government project for machine translation. *Computational Linguistics* **11**(2–3): 91–110.

Nenkova, A., and McKeown, K. 2012. A survey of text summarization techniques. In C. C. Aggarwal and C. X. Zhai (eds.), *Mining Text Data*. Boston, MA: Springer, pp. 43–76.

Oliveira, H., Ferreira, R., Lima, R., Lins, R.F., Freitas, F., Riss, M., and Simske, S.J. 2016. Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. *Expert Systems with Applications* **65**: 68–86.

Olvera-Lobo, M.D., and Gutierrez-Artacho, J. 2015. Question answering track evaluation in TREC, CLEF and NTCIR. In A. Rocha, A. Correia, S. Costanzo, and L. Reis (eds.), *New Contributions in Information Systems and Technologies - Advances in Intelligent Systems and Computing*, p. 353, Berlin, Germany: Springer.

Pechsiri, C. and Piriyakul, R. 2016. Developing a why-how question answering system on community web boards with a causality graph including procedural knowledge. *Information Processing in Agriculture* **3**: 36–53.

Qiang, J.-P., Chen, P., Ding, W., Xie, F., and Wu, X. 2016. Multi-document summarization using closed patterns. *Knowledge-Based Systems* **99**: 28–38.

Sak, H., Güngör, T., and Saraçlar, M. 2011. Resources for Turkish morphological processing. *Language Resources and Evaluation* **45**: 249–61.

Shekarpour, S., Marx, E., Ngomo, A.-C.N., and Auer, S. 2015. SINA: semantic interpretation of user queries for question answering on interlinked data. *Web Semantics: Science, Services and Agents on the World Wide Web* **30**: 39–51.

Silber, H.G., and McCoy, K.F. 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics* **28**(4): 487–96.

Utomo, F.S., Suryana, N., and Azmi, M.S. 2017. Question answering system: a review on question analysis, document processing, and answer extraction techniques. *Journal of Theoretical and Applied Information Technology* **95**(14): 3158–74.

Wan, X. 2009. Topic analysis for topic-focused multi-document summarization. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM*, pp. 1609–12.

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141

Wang, D., and Nyberg, E. 2015. A long short-term memory model for answer sentence selection in question answering. In *Proceedings of ACL-IJCNLP*, pp. 707–12.

Wang, D., Zhu, S., Li, T., and Gong, Y. 2012. Comparative document summarization via discriminative sentence selection. *ACM Transactions on Knowledge Discovery from Data* **6**(3), 12:1–12:18.

Wu, Y., Hori, C., Kashioka, H., and Kawai, H. 2015. Leveraging social Q&A collections for improving complex question answering. *Computer Speech and Language* **29**: 1–19.

Xiong, S., and Ji, D. 2016. Query-focused multi-document summarization using hypergraph-based ranking. *Information Processing and Management* **52**: 670–81.

Xiong, C., Merity, S., and Socher, R. 2016. Dynamic memory networks for visual and textual question answering. In *Proceedings of the International Conference on Machine Learning*, pp. 2397–406.

Yang, L., Ai, Q., Spina, D., Chen, R-C., Pang, L., Croft, W.B., Guo, J., and Scholer, F. 2016. Beyond factoid QA: effective methods for non-factoid answer sentence retrieval. In *Proceedings of the European Conference on Information Retrieval (ECIR*, pp. 115–28.

Yang, M.-C., Lee, D.-G., Park, S.-Y., and Rim, H.-C. 2015a. Knowledge-based question answering using the semantic embedding space. *Expert Systems with Applications* **42**: 9086–104.

Yang, Y., Yih, W.-t., and Meek, C. 2015b. WIKIQA: A challenge dataset for open-domain question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2013–18.

Yih, W-T., He, X., and Meek, C. 2014. Semantic parsing for single-relation question answering. In *Proceedings of the Annual Meeting of ACL*, pp. 643–8.

Yu, L., Hermann, K.M., Blunsom, P., and Pulman, S. 2014. Deep learning for answer sentence selection, In *Proceedings of NIPS Deep Learning Workshop.*

Zheng, Z. 2002. AnswerBus question answering system. In *Proceedings of the International Conference on Human Language Technology Research (HLT*, pp. 399–404.

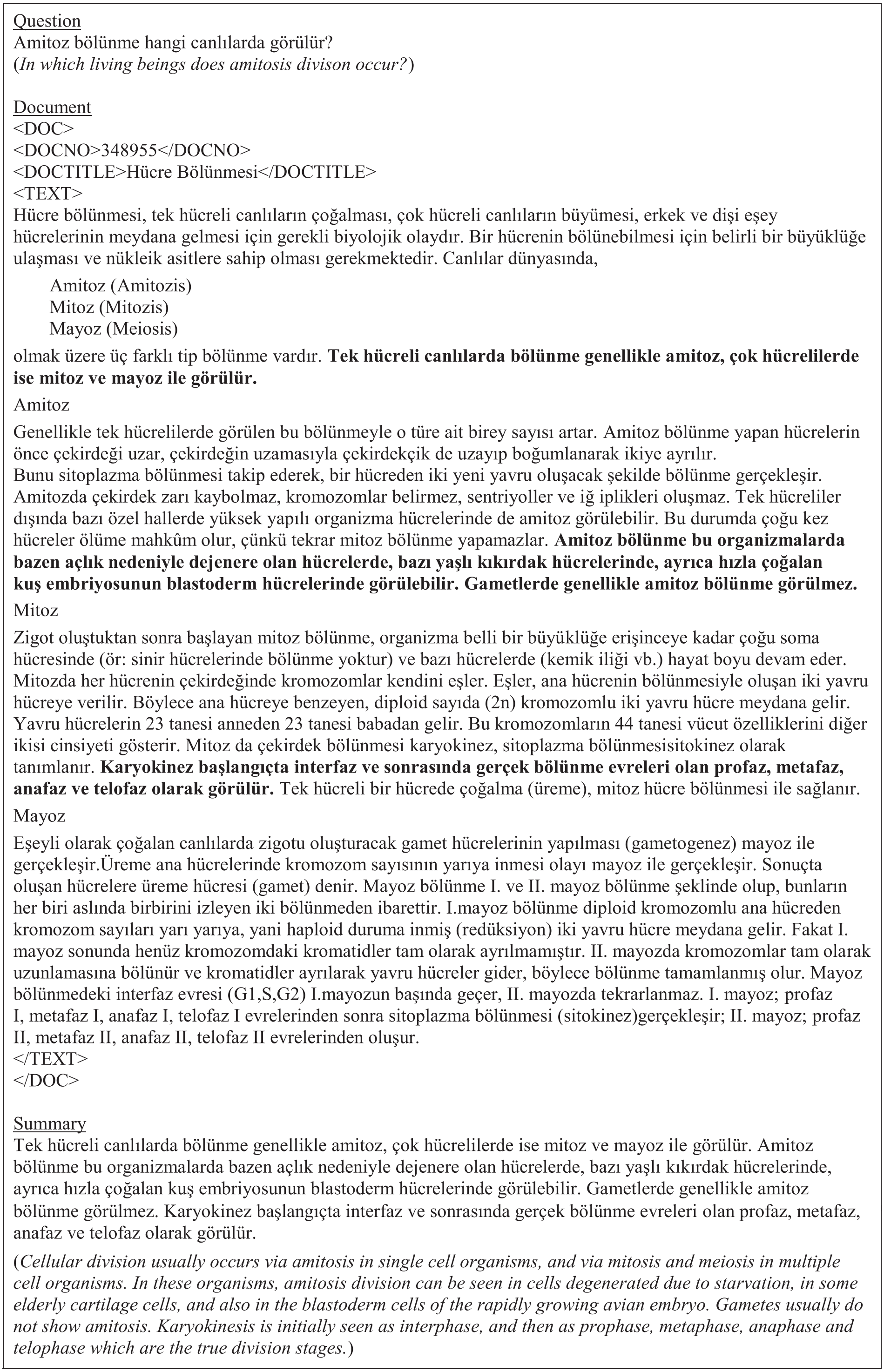
Zhong, S.-h., Liu, Y., Li, B., and Long, J. 2015. Query-oriented unsupervised multi-document summarization via deep learning model. *Expert Systems with Applications* **42**: 8146–55.

Appendix A. Example for the summarization process

An example question, a document (titled “hücre bölünmesi” (“cell division”)) returned by the system and the summary of the document are shown in Figure A1. The summary sentences are shown in bold within the document and also the summary is given below the document. The first sentence in the summary includes the answer to the question.

The effect of the three summarization features can be seen in this example summary. The words in the question (e.g., “amitoz” (amitosis), “bölünme” (division)) have an important effect in the process; we can see these words in most of the summary sentences. Some of the words (e.g., “hücre” (cell)) have a high-term frequency score and they increase the scores of the sentences including them. Also, the lexical chain corresponding to the document contributed to the scores of some of the sentences in the document. For instance, the words “metafaz” (metaphase), “anafaz” (anaphase), and “telofaz” (telophase) that appear in one of the summary sentences are related to the word “mayoz” (meiosis) in the Biology ontology. The word “mayoz” (meiosis) is a frequent term in the document and this caused its related terms to be included in the lexical chain. Thus, each of these related terms in the sentence caused the sentence score to be increased.

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141



Question
Amitoz bölünme hangi canlılarda görülür?
(*In which living beings does amitosis divison occur?*)

Document
<DOC>
<DOCNO>348955</DOCNO>
<DOCTITLE>Hücre Bölünmesi</DOCTITLE>
<TEXT>
Hücre bölünmesi, tek hücreli canlıların çoğalması, çok hücreli canlıların büyümesi, erkek ve dişi eşey hücrelerinin meydana gelmesi için gerekli biyolojik olaydır. Bir hücrenin bölünebilmesi için belirli bir büyüklüğe ulaşması ve nükleik asitlere sahip olması gerekmektedir. Canlılar dünyasında,

Amitoz (Amitozis)
Mitoz (Mitozis)
Mayoz (Meiosis)

olmak üzere üç farklı tip bölünme vardır. **Tek hücreli canlılarda bölünme genellikle amitoz, çok hücrelilerde ise mitoz ve mayoz ile görülür.**

Amitoz

Genellikle tek hücrelilerde görülen bu bölünmeyle o türe ait birey sayısı artar. Amitoz bölünme yapan hücrelerin önce çekirdeği uzar, çekirdeğin uzamasıyla çekirdekçik de uzayıp boğumlanarak ikiye ayrılır.
Bunu sitoplazma bölünmesi takip ederek, bir hücreden iki yeni yavru oluşacak şekilde bölünme gerçekleşir. Amitozda çekirdek zarı kaybolmaz, kromozomlar belirmez, sentriyoller ve iğ iplikleri oluşmaz. Tek hücreliler dışında bazı özel hallerde yüksek yapılı organizma hücrelerinde de amitoz görülebilir. Bu durumda çoğu kez hücreler ölüme mahkûm olur, çünkü tekrar mitoz bölünme yapamazlar. **Amitoz bölünme bu organizmalarda bazen açlık nedeniyle dejenere olan hücrelerde, bazı yaşlı kıkırdak hücrelerinde, ayrıca hızla çoğalan kuş embriyosunun blastoderm hücrelerinde görülebilir. Gametlerde genellikle amitoz bölünme görülmez.**

Mitoz

Zigot oluştuktan sonra başlayan mitoz bölünme, organizma belli bir büyüklüğe erişinceye kadar çoğu soma hücresinde (ör: sinir hücrelerinde bölünme yoktur) ve bazı hücrelerde (kemik iliği vb.) hayat boyu devam eder. Mitozda her hücrenin çekirdeğinde kromozomlar kendini eşler. Eşler, ana hücrenin bölünmesiyle oluşan iki yavru hücreye verilir. Böylece ana hücreye benzeyen, diploid sayıda (2n) kromozomlu iki yavru hücre meydana gelir. Yavru hücrelerin 23 tanesi anneden 23 tanesi babadan gelir. Bu kromozomların 44 tanesi vücut özelliklerini diğer ikisi cinsiyeti gösterir. Mitoz da çekirdek bölünmesi karyokinez, sitoplazma bölünmesisitokinez olarak tanımlanır. **Karyokinez başlangıçta interfaz ve sonrasında gerçek bölünme evreleri olan profaz, metafaz, anafaz ve telofaz olarak görülür.** Tek hücreli bir hücrede çoğalma (üreme), mitoz hücre bölünmesi ile sağlanır.

Mayoz

Eşeyli olarak çoğalan canlılarda zigotu oluşturacak gamet hücrelerinin yapılması (gametogenez) mayoz ile gerçekleşir.Üreme ana hücrelerinde kromozom sayısının yarıya inmesi olayı mayoz ile gerçekleşir. Sonuçta oluşan hücrelere üreme hücresi (gamet) denir. Mayoz bölünme I. ve II. mayoz bölünme şeklinde olup, bunların her biri aslında birbirini izleyen iki bölünmeden ibarettir. I.mayoz bölünme diploid kromozomlu ana hücreden kromozom sayıları yarı yarıya, yani haploid duruma inmiş (redüksiyon) iki yavru hücre meydana gelir. Fakat I. mayoz sonunda henüz kromozomdaki kromatidler tam olarak ayrılmamıştır. II. mayozda kromozomlar tam olarak uzunlamasına bölünür ve kromatidler ayrılarak yavru hücreler gider, böylece bölünme tamamlanmış olur. Mayoz bölünmedeki interfaz evresi (G1,S,G2) I.mayozun başında geçer, II. mayozda tekrarlanmaz. I. mayoz; profaz I, metafaz I, anafaz I, telofaz I evrelerinden sonra sitoplazma bölünmesi (sitokinez)gerçekleşir; II. mayoz; profaz II, metafaz II, anafaz II, telofaz II evrelerinden oluşur.
</TEXT>
</DOC>

Summary
Tek hücreli canlılarda bölünme genellikle amitoz, çok hücrelilerde ise mitoz ve mayoz ile görülür. Amitoz bölünme bu organizmalarda bazen açlık nedeniyle dejenere olan hücrelerde, bazı yaşlı kıkırdak hücrelerinde, ayrıca hızla çoğalan kuş embriyosunun blastoderm hücrelerinde görülebilir. Gametlerde genellikle amitoz bölünme görülmez. Karyokinez başlangıçta interfaz ve sonrasında gerçek bölünme evreleri olan profaz, metafaz, anafaz ve telofaz olarak görülür.

(*Cellular division usually occurs via amitosis in single cell organisms, and via mitosis and meiosis in multiple cell organisms. In these organisms, amitosis division can be seen in cells degenerated due to starvation, in some elderly cartilage cells, and also in the blastoderm cells of the rapidly growing avian embryo. Gametes usually do not show amitosis. Karyokinesis is initially seen as interphase, and then as prophase, metaphase, anaphase and telophase which are the true division stages.*)

Fig. A1. An example question, document and summary.

Downloaded from https://www.cambridge.org/core. Bogazici Universitesi, on 09 Oct 2018 at 08:12:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324918000141