A Novel Question Answering System for Albanian Language

Evis Trandafili(✉), Elinda Kajo Meçe, Kristjan Kica, and Hakik Paci

Department of Computer Engineering, Faculty of Information Technology, Polytechnic University of Tirana, Tirana, Albania
{etrandafili, ekajo, kristjan.kica, hpaci}@fti.edu.al

Abstract. The volume of unstructured data is constantly growing, drawing the attention of the research community toward Natural Language Processing tasks. Recent advances in Information Extraction have led to the implementation of different systems and tools for Question Answering. These approaches are mainly language dependent as they need information about the language structure and syntax to perform well. This paper proposes an approach to extracting answers of factoid questions for a given text in Albanian Language. As far as we know, this is the first attempt of a Question Answering system for Albanian language. Experiments show that this is an effective solution for single domain documents.

1 Introduction

The volume of available digital information is constantly growing and the way we live and work relies more and more on the universal information available. The freedom of publishing has led to replicated information usually mixed with other non-crucial data. Moreover, this large volume of information is mainly unstructured thus finding relevant information becomes more complicated.

The main objective of a question answering system is to provide the exact required information with much less human efforts. The users of digital information tend to query data in natural language. Traditional information retrieval systems respond to a query with a list of the most relevant documents where the user has to investigate for the required answer. This is time consuming. From a QA system point of view, information retrieval techniques are used to extract the exact information within the document which responds to the question. Most of question answering systems are implemented based on factoid questions. These are the type of questions whose answers are simple facts expressed with a short string which refers to a date, a place, a name, etc. [1]. For example, questions like *"Kur u shpall pavarësia e Shqipërisë?", (English: When was Albania proclaimed independent?), "Ku buron lumi Shkumbin?", (English: Where does the Shkumbin river originate?), "Kush e shkroi librin Meshari?", (English: Who wrote the book Meshari?),* correspond to some factoid questions in Albanian language.

A typical question answering system first determines the answer type by processing the query/question and then reformulates it in the appropriate query format of the

L. Barolli et al. (Eds.): EIDWT 2018, LNDECT 17, pp. 514–524, 2018.
https://doi.org/10.1007/978-3-319-75928-9_46

search engine. The outputs are the relevant ranked documents broken into passages. Finally, text answers are extracted and ranked. Our QA system is based on the above method. Language dependent rules are implemented to detect the type of factoid question and to highlight the answer type. We used Text-Based question answering paradigms to extract the answer type and formulate the query relying on the lexical and semantic matching of the key words present both in the question and in the document.

The organization of the paper is as follows: Sect. 2 presents the background and the related works on question answering systems; Sect. 3 presents some basic information about the Albanian language structure; Sect. 4 presents the architecture and some implementation details of our novel QA system; Sect. 5 discusses the evaluation metrics used to test the performance of our system; and we conclude in Sect. 6 with conclusions and future work.

2 Background and Related Works

Question answering is a subfield of Information Retrieval and Natural Language Processing which focuses on the extraction of passages within the document in response to the user's need for information expressed in natural language. Ongoing efforts are done to automate and improve this process. There are three approaches for implementing Question Answering systems; the simplest is the IR-based model which focuses on factoid questions; knowledge-based model relies on the semantic representation of the query; and the hybrid model uses text corpuses and structured knowledge databases [1]. Most of QA systems are based on factoid questions due to the implementation simplicity, whereas knowledge-based approaches aim to answer questions about definitions or concepts and are much harder to implement.

The standard algorithm for QA based on factoid questions encompasses three main modules: the question processing module which extracts the answer type (identifying the entity of the answer) and the keywords for the IR system to retrieve passages in the document. Next, the passage retrieval module uses an answer type classification to filter out passages that contain the wrong answers and then rank the remaining passages using supervised machine learning algorithms. The final module is answer extraction where the extraction process is achieved using information about the expected answer type together with regular expression patterns or using N-gram tiling when QA is applied in web search [1].

Efforts in the implementation of QA systems have been made since 1961 with BASEBALL [2] that answered questions about baseball games and LUNAR [3] that provided answers about soil samples taken from Apollo exploration. Both these systems answered English written questions by transforming the question into a database query through pattern matching rules.

Ongoing attempts aim to improve the overall performance of QA systems. Watson is based on a parallel framework composed of two modules: Natural Language Processing module which deals with question analysis and Information Extraction module which retrieves the candidate answers [4].

Another approach on QA uses the star architecture by viewing the subtasks (question processing, passage retrieval and answer extraction) as nodes connected by a

central node that generates the optimal strategy to find the answer depending on the type of question [5].

An interesting approach is a factoid-based QA system named Sybil, which works with spoken documents in English [6]. It uses natural language analyzers and linguistic information obtained with machine learning tools. Sybil was evaluated using the European Parliament Plenary Sessions English corpus and its performance is better than the state-of-the-art on this corpus.

A QA system which finds the answers from a single document based on the category space acquired from Wikipedia is proposed in [7]. A Natural Language Processing Toolkit is used for keyword extraction and the distance between categories is used to rank the answers.

There are also several attempts to connect different domains of artificial intelligence with question answering. The authors in [8] focused on the state of the art of Visual Question Answering, a field of study that combines Computer Vision with Natural Language Processing methods. They reviewed different approaches that aim to map questions and images in vector representation using a common feature space. They highlighted that a successful approach is the combination of convolutional neural networks that are trained on object recognition, with word embedding's, trained on large text corpora.

A new way of thinking regarding question answering is proposed in [9]. The proposed architecture automatically generates questions from sentences containing important facts or events and their respective answers. This QA system builds and maintains a database of pairs (question, answer) corresponding to the domain of documents. The main components used for the system are: sentence split, named-entity recognition, question generation, question filtering and question/answer indexing. The user is presented with a set of candidate query questions for his information needs.

Moreover, substantial works are made to handle questions in other than English languages. A QA monolingual system for searching French documents based on French questions is proposed by [10]. They used a named entity recognition technique and a syntactic analyzer to identify the candidate answers. Following, a matching strategy is applied to rank the answers. The bilingual module is able to answer questions written in Dutch, German, Italian, Portuguese, Spanish, English and Bulgarian. This is achieved by automatically translating the original question into French and then proceeding with the monolingual QA system.

Moreover, a bilingual question answering system is proposed in [11] and handles Bangla and English electronic documents. The authors claim that this system generates questions and answers efficiently. It has four main components: database initialization and processing, storage, question answering and query execution.

Several efforts are made to support questions made in Arabic language. A QA system for Arabic language is implemented in [12]. This system handles different types of questions and relies on a language dependent preprocessing step to increase its overall performance.

Furthermore, in [13] the authors introduce a QA system for Vietnamese language based on ontology. The system has two modules: the question analysis module and the answer retrieval module. The semantic structure of the question is captured through an intermediate representation and then it is matched with the target ontology.

3 Albanian Language Structure

Albanian language is considered a separate branch of the Indo-European language family and cannot conclusively be closely connected with any other Indo-European language. It is the official language of Albania, the co-official language of Kosovo, and the co-official language of many western municipalities of the Republic of Macedonia. Albanian is also spoken widely in some areas in Greece, southern Montenegro, southern Serbia, and in some towns in southern Italy and Sicily.

The Albanian language has two main dialects spoken in two major regions; the north dialect called Gheg and the southern dialect called Tosk. The official Albanian language is written in Roman alphabet and from 1909 till World War II was based on the south Gheg dialect. Since then it has been modeled on Tosk dialect [14]. For the purpose of this paper, the official Albanian language is used. It has 7 vowels and 29 consonants. The vowels are represented by single Latin letters (*a, e, ë, i, o, u, y*), and the consonants by single letters (*b, c, ç, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, x, z*), and combination of different letters (*dh, gj, ll, nj, rr, sh, th, xh, zh*).

Even though the most common word order in a sentence is Subject-Verb-Object (SVO), Albanian like German is relatively free [15] because the role of the word in a sentence is not determined by its occurring position, but by its inflectional ending and by its relative meaning. For example, the sentence *"Teuta read all the documents."* can have different possible role orders with only slight pragmatic differences. All the sentences listed below are grammatically correct.

SVO - *Teuta i lexoi të gjitha dokumentat.*
SOV - *Teuta, të gjitha dokumentat i lexoi.*
VOS - *I lexoi të gjitha dokumentat Teuta.*
VSO - *I lexoi Teuta të gjitha dokumentat.*
OVS - *Të gjitha dokumentat i lexoi Teuta.*
OSV - *Të gjitha dokumentat Teuta i lexoi.*

Albanian questions are linguistic expressions used to compose a demand for information. In this paper we focus only on factoid questions, which are questions that can be answered with simple facts expressed in short text answers [1]. One of the main attributes of the syntactic structure of factoid questions are the interrogative pronouns and adverbs that are also known as wh-words or function words. In Albanian language the most common wh-words are: *kush – who; cili – which; çfarë/çka/ç'/se – what; pse/përse – why; nga/ku - where*. There are other question words such as *si/qysh – how*; *sa – how much; i/e satë/i/a* asks for the order of something; *sejtë* asks about the composition of something; *kush* can be associated with a preposition and has the grammatical function of case; *cili* and *satë* have the grammatical function of case, number and gender [16].

4 Algorithmic Approach and System Design

In this section we introduce the design and the implementation details of the question answering system.

4.1 System Architecture

Our novel question answering system is composed by three main modules:

1. Document preprocessing and indexing module that encompasses language-dependent rules.
2. Question analysis module that extracts the answer type and generates the query taken as input by the next module.
3. Passage Retrieval and Answer Extraction module finds the candidate answers using the information catalogued in the preprocessing module, the query generated by the question analysis as well as additional information from an external database of language dependent data.

The above modules are implemented as a java desktop application (Fig. 1).

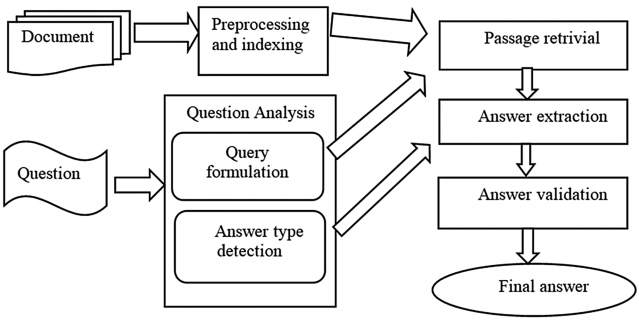

Fig. 1. Architecture of QA system for Albanian language

4.2 General Assumptions and Additional Resources

The proposed QA system is based on the assumption that the questions are factoid and somehow related to a single predefined document. Also we assume that the questions are correctly written in standard official Albanian language. Even though the system may answer correctly, it performs better when user types the letters *ë* and *ç*, instead of using *e* and *c*. Furthermore, our system aims to extract answers from a single predefined document.

4.3 Document Preprocessing

The text document is preprocessed beforehand. We extended Albanian abbreviations to their original word form. For example we converted *shek*. to *shekull,* (English : *century*), *dt*. to *datë,* (English : *date*), *etj*. to *e të tjerë,* (English : *etcetera*), *p.e.s.* to *para erës sonë,* (English : *Before the Common Era*) and *e.r.* to *era e re*, (English : *common era*). After that, named entities are extracted by hand and indexed for later faster searches. For the purpose of this paper, the named entities taken in consideration are: PERSON, LOCATION, DATE, TIME, REASON, NUMBER, MANNER, LANGUAGE and PUBLICATION. Furthermore, the document is preprocessed using

tokenization, stop-words/common verbs removal and stemming algorithms [17]. As a result, the final text is cleaned from the inflectional endings and suffixes allowing a better word to word matching with the keywords extracted from the question. The previous mentioned algorithms are implemented in java programming language and are based on several rules comprising the Albanian language structure.

The document pre-processing algorithm:

```
Abbreviations' extension
NER
Tokenization
Stop-word and common verb removal
Stemming
    remove_inflectional_endings
    remove_sufixes
```

4.4 Question Analysis

This module includes answer type detection and query formulation. The answer type will be used to extract from the document the name entity that corresponds to the requested answer while the generated query will contribute in the extraction of passages that are most likely to contain the answer.

The most important attribute that we use to extract the answer type, is the wh-word. Since the standard question structure in Albanian is $<preposition> + [question\ word] + [\ldots]$ we make use of the following wh-words: "*kush, ku, kur, pse, sa*", that gives us enough information to extract the answer types respectively, PERSON, LOCATION, TIME/DATE, REASON, NUMBER. If a more general wh-word is identified, such as *çfarë, ç', cili* we utilize the question headword which is the first noun phrase after the verb. To facilitate this task, we constructed a database of headwords that map to their respective answer type as shown in Table 1.

Table 1. Overview of the data collected to map headwords with their answer type.

Headword	Answer type
komandant, sulltan, djalë,vajzë (English: commander, sultan, boy, girl)	PERSON
vise, tokë, qytet, shtet, principatë (English: land, land, city, state, principality)	PLACE/LOCATION
datë, vit (English: date, year)	DATE/TIME/YEAR
mënyrë, metodë (English: way, method)	MANNER
…	…

The headwords give us additional information beyond the specific wh-words. For example: *"Në cilin vit u shkrua 'Historia dhe gjenealogjia e shtepisë së Muzakajve'?" (English: In which year was the 'Story and genealogy of the Muzaka family' written?)*, tells us that the user is asking for a YEAR, which is more specific than the wh-word

"*kur*" which only tells us that a TIME entity is required. If a supplementary entity is not found, TIME entity is provided instead.

In case the wh-word is not located in the beginning but somewhere inside the question string, we also search for it inside the user question. If still the entity is not found, we search through the question for keywords associated with certain answer types. For example: *"Çfarë ka shkruar Marin Barleti për jetën e Skënderbeut?", (English: What has Marin Barleti written about Skanderbeg's life?),* using *shkruar* we can deduce that the user is asking about a PUBLICATION.

Furthermore, the question words are removed from the question string and the same preprocessing algorithms applied to the text document are also applied to the question string.

The answer type detection algorithm:

```
answer_type=null
Find the question_word
If general question_word
  Find headword
  If headword found and headword in database
    answer_type=get_answer_type_from_headword
  else
    Find words associated with answer types
    If associated_words
      answer_type=get_answer_type_from_associated_words
    else
      answer_type=default_from_general_question_word
else
  answer_type=answer_type_from_specific_question_word
```

4.5 Passage Retrieval and Answer Extraction

The tdf/if representation is generated and the cosine similarity between the query string and the text passages is calculated. Based on cosine similarity, we ranked the document passages by the probability of containing the correct answer. To account for the common user input mistakes with the letters *ë* and *ç*, we replace them with *e* and *c*.

At this stage, we have a list of previously ranked text passages. From the first 5 candidate answer passages, we extract the entities that comply with the answer type entity extracted previously. Since the search is being performed in a single document, the answer is unlikely expressed more than once. Instead of the lexical matching between the query and the passages we also expanded it with synonyms, homonyms and hyponyms [18] to account for different ways of asking the same question. These word extensions are maintained in a database constructed for the purpose of expanding the query words.

The candidate answers are ranked using the following criteria:

1. Cosine Similarity with idf between the passage containing the requested entity and the query string.
2. Number of query words in the candidate answer passage.

3. Average word distance between the query words that are present in the passage and the candidate entity.
4. Longest sequence of query words in the candidate passage.
5. Number of query words less than 4 words apart from the candidate in the passage.

We used the metrics discussed previously to evaluate the rank value. The candidate with the higher value assigned is most likely to be the right answer. The formula we used to calculate the final evaluation is as follows:

$$E = \text{nr\_of\_words} * \text{cos\_similatity} * 100 * (1 + \text{longest\_sequence}/5) * \text{proximity}. \quad (1)$$

where proximity is the inverse of the calculated value for average word distance.

The ranking algorithm applied to the candidate answers is illustrated in Tables 2 and 3 with two different example questions.

Table 2. Question 1- *Kur lindi Skënderbeu? (English: When was Skanderbeg born?)*

Candidate	në vitin 1405	në vitin 1468
Context	*Skënderbeu ose Skënderbej lindi në vitin 1405 dhe vdiq në vitin 1468 (English: Skanderbeg was born in 1405 and died in 1468)*	
Proximity	3.83	1.64
Number of words	2	2
Longest sequence	2	2
Cosine similarity	0.062	0.062
Evaluation	65.90	15.69

The evaluation is calculated based on the values assigned to proximity, number of words, longest sequence and cosine similarity. The answer is the candidate with the highest evaluation. In the example in Table 2, the evaluation is mainly influenced by the value of proximity, whereas in the example shown in Table 3 the ranking is more difficult.

5 Performance Evaluation

To evaluate our novel question answering system, we used a text document in Albanian language with 1300 words, regarding *Skanderbeg/Gjergj Kastrioti*, the Albanian national hero.

The first metric of evaluation used is the percentage of correct answers where we considered the candidates ranked first against the correct answers manually labeled. We use the Precision metric as in to the following formula:

Table 3. Question 2- *Sa motra kishte Skënderbeu? (English: How many sisters did Skanderbeg have?)*

Candidate	5	4	12 vjet
Context	*Gjergj Kastrioti ishte djali më i vogël i Gjon Kastriotit dhe i princeshës Vojsava, fëmija i fundit midis 4 djemve dhe 5 vajzave. (English: Gjergj Kastrioti was the youngest son of Gjon Kastriot and Princess Vojsava, the last child between four boys and five girls)*		*Shkrimet latinisht të Frangut të vitit 1480, 12 vjet pas vdekjes së Skënderbeut, mjerisht u përvetësuan nga të tjerë, dhe përkthimi dhe botimi i saj italisht u bë më vonë, pas vdekjes. (English: The Latin script of Frang in 1480, 12 years after Scanderbeg's death, was unfortunately taken by others, and its translation and publication to Italian was done after his death)*
Proximity	0.34	0.27	0.31
Number of words	0.70	0.70	1.00
Longest sequence	0.70	0.70	1.00
Cosine similarity	0.10	0.10	0.0064
Evaluation	2.78	2.18	0.24

$$\text{Precision} = \frac{\text{m}}{\text{n}} * 100\%. \tag{2}$$

In (2), n is the total number of questions, and m is the number of correctly answered questions.

We also applied the Mean Reciprocal rank (3) to the ranked list of candidates provided by the system. Each question is scored as the inverse of the first correct answer in the list of candidates. The following formula is used:

$$\text{MRR} = \frac{1}{\text{n}} * \sum_{i=1}^{n} \frac{1}{j}. \tag{3}$$

In (3), n is the total number of questions and j is the rank of the correct answer in the ranked list previously generated by the QA system.

Out of 138 questions, we achieved a precision value of 69.5%, and a mean reciprocal rank scoring 73.5%.

6 Conclusions and Future Work

In this paper we introduced a novel question answering system which can answer to factoid questions based on a single text document written in Albanian language. As far as we know there are no previous works done in QA for Albanian language. We implemented a prototype of a QA system that is composed by three main modules:

Document preprocessing; Question analysis; Passage Retrieval and Answer Extraction. The majority of language dependent tasks are comprised in the preprocessing module even if some language dependent rules are also needed inside the question analysis module for identifying the answer type. Furthermore, we used the tf/idf representation of the document and calculated the cosine similarity to identify the similarity between the query string and the text passages.

We evaluated the QA system using a specific domain and by asking questions regarding that domain. For the evaluation process we used two different statistical metrics and achieved a precision value of 69.5% and a mean reciprocal rank value of 73.5%.

As a future work, our novel QA system can be further automated by using Machine Learning approaches for Named Entity Recognition in Albanian language. Some humble works are previously done in this context but it does not exist a formal tool that can be applied in different NLP tasks.

Moreover, the text preprocessing module should be enhanced with new grammatical and syntactic rules. An extended database of synonyms, homonyms and hypernyms should be comprised to support multiple domain questions. The performance of the QA system can also benefit from the incorporation of a Part of Speech tagger in the preprocessing module.

Furthermore, the question analysis module can be improved by using Machine Learning classification algorithms to best determine the answer type and to improve the system's precision.

As this QA system considers only factoid questions, additional efforts should be done to extend the range of questions with an alternative list of questions.

References

1. Jurafsky, D., Martin, J.H.: Speech and Language Processing, 2nd edn. Prentice-Hall, Inc., Upper Saddle River, New Jersey (2009). ISBN 0131873210
2. Green, B.F., Wolf, A.K., Chomsky, C., Laughery, K.: Baseball: an automatic question answerer. In: Proceedings of the Western Joint Computer Conference, vol. 19, pp. 219–224 (1961). Reprinted in Grosz et al. (1986)
3. Woods, W.: Progress in natural language understanding - an application to lunar geology. In: Proceedings of AFIPS Conference, vol. 42, pp. 441–450 (1973)
4. Ferrucci, D.A.: Introduction to "this is Watson". IBM J. Res. Develop. **56**(3/4), 1–15 (2012)
5. Nyberg, E., et al.: The JAVELIN question answering system at TREC2003: a multi-strategy approach with dynamic planning. In: The Proceedings of the 11th Text Retrieval Conference (2003)
6. Comas, P.R., Turmo, J., Màrquez, L.: Sibyl a factoid question-answering system for spoken documents. ACM Trans. Inf. Syst. **30**(3), Article 19, 40 (2012)
7. Wang, X., Xu, B., Zhuge, H.: Automatic question answering based on single document. In: 12th International Conference on Semantics Knowledge and Grids. IEEE (2016)
8. Wu, Q., et al.: Visual question answering: a survey of methods and datasets. Comput. Vis. Image Underst. **163**, 21–40 (2017). https://doi.org/10.1016/j.cviu.2017.05.001

9. Kim, M., Kim, H.: Design of question answering system with automated question generation. In: Fourth International Conference on Networked Computing and Advanced Information Management. IEEE (2008)
10. Perret, L.: A question answering system for French. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) Multilingual Information Access for Text, Speech and Images. CLEF 2004. Lecture Notes in Computer Science, vol 3491. Springer, Berlin, Heidelberg (2005)
11. Hoque, S., et al.: BQAS: a bilingual question answering system. In: Proceedings of International Conference on Electrical Information and Communication Technology. IEEE (2015). https://doi.org/10.1109/eict.2015.7392020
12. Kamal, A.I.: Enhanced Arabic question answering system. In: Sixth International Conference on Computational Intelligence and Communication Networks. IEEE (2014)
13. Nguyen, D.Q., et al.: A vietnamese question answering system. In: International Conference on Knowledge and Systems Engineering. IEEE (2009)
14. Hamp, E.P.: Albanian language, Encyclopedia Britannica (2016)
15. Kurani, A., Trifoni, A.: Syntactic similarites and differences between Albanian and English. Eur. Sci. J. **17** (2011)
16. Çabej, E.: Studime etimologjike në fushë të shqipes I, Tiranë (1982)
17. Sadiku, J., Biba, M.: Automatic stemming of Albanian through a rule-based approach. J. Int. Res. Publ. Lang. Individuals Soc. **6** (2012). ISSN-1313-2547
18. Dhrimo, A., Tupja, E., Ymeri, E.: Fjalor sinonimik i gjuhës shqipe: Mbi 30 000 zëra. Botimet Toena, Tiranë (2002)