A Comparative Evaluation of QA Systems over List Questions

Patricia Nunes Gonçalves and António Horta Branco(✉)

Department of Informatics, University of Lisbon, Edifício C6,
Faculdade de Ciências Campo Grande, 1749-016 Lisbon, Portugal
{patricia.nunes,antonio.branco}@di.fc.ul.pt

Abstract. The evaluation of a Question Answering system is a challenging task. In this paper we evaluate our system, LX-ListQuestion, a Web-based QA System that focuses on answering list questions. We compare our system against other QA Systems and the results were analyzed in two ways: (i) the quantitative evaluation of answers provides recall, precision and F-measure and (ii) the question coverage that indicate the usefulness of the system to the user by counting the number of questions for which the system provides at least one correct answer. The evaluation brings interesting results that points to a certain degree of complementary between different approaches.

Keywords: QA Systems · List questions · Evaluation QA

1 Introduction

In Open-domain Question Answering the range of possible questions is not constrained, hence a much tougher challenge is placed on systems. The goal of an Open-domain QA system is to answer questions on any kind of subject domain [10]. Research in Open-domain Question Answering had a boost in 1999 with the Text REtrieval Conference (TREC)[1], which provides large-scale evaluation of QA systems thus defining the direction of research in the QA field.

List questions started being studied in the context of QA in 2001 when TREC included this type of questions in the dataset.

Finding the correct answers to List questions requires discovering a set of different answers in a single document or across several documents. An approach to answer a List question in a single document is very similar to the approach to find the correct answer to factoid questions: (i) find the most relevant document; (ii) find the most relevant excerpt and (iii) extract the answers from this relevant excerpt. On the other hand, the process to extract the answers spread over several documents raised new challenges such as grouping repeated elements, handling more information, separating the relevant information from the rest of the information, among others.

[1] http://trec.nist.gov/.

J. Silva et al. (Eds.): PROPOR 2016, LNAI 9727, pp. 115–121, 2016.
DOI: 10.1007/978-3-319-41552-9_11

Evaluation of QA Systems involves a large amount of manual effort, but it is a fundamental component to improve the systems. Traditional evaluation of QA systems use recall, precision and F-measure to measure performance of systems [8,11]. Besides the traditional evaluation, we assessed the systems by using question coverage that indicate the usefulness of the system to the user by counting the number of questions that the system provides at least one correct answer, providing another perspective of evaluation. **Paper Outline**: Sect. 2 introduces our system, LX-ListQuestion, a Web-based QA system that uses redundancy and heuristics to answer List questions. In Sect. 3 we compare the results, over the same question dataset, with other two QA systems: RapPortagico and XisQuê. Finally in Sect. 4 we present some concluding remarks.

2 LX-ListQuestion Question Answering System Architecture

The LX-ListQuestion System [6,7] is a fully-fledged Open-domain Web-based QA system for List questions. The system collect answers spread over multiple documents using the Web as a corpus. Our approach is based on redundancy of information available on the Web combined with heuristics to improve QA performance. The implementation is guided by the following design features:

- Exploits redundancy to find answers to List questions;
- Compiles and extracts the answers from multiple documents;
- Collects at run-time the documents from Web using a search engine;
- Provides answers in real time without resorting to previously stored information.

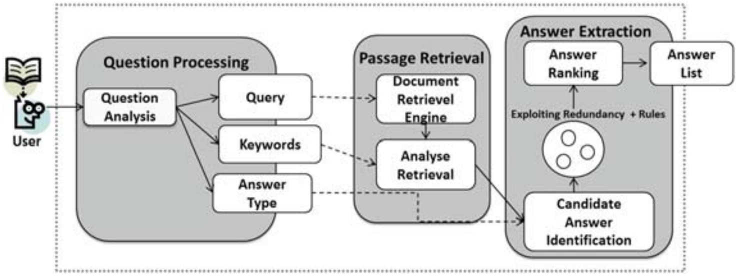

Fig. 1. LX-ListQuestion system architecture

The system architecture is composed by three main modules: Question Processing, Passage Retrieval and Answer Extraction (Fig. 1). The Question Processing module is responsible for converting a natural language question into a form that subsequent modules are capable of handling. The main sub-tasks are (i) question analysis: responsible for cleaning the questions; (ii) extraction of keywords: performed using nominal expansion and verbal expansion; (iii) transformation of the question into a query; (iv) identification of the semantic category of the expected answer; and (v) identification of the question-focus.

The Passage Retrieval Module is responsible for searching Web pages and saving their full textual content into local files for processing. After the content is retrieved, the system will select relevant sentences. The Answer Extraction Module aims at identifying and extracting relevant answers and presenting them in list form. The candidate answer identification is based on a Named Entity Recognition tool. The candidates are selected if they match the semantic category of the question. The process of building the Final List Answer and more details about implementation of LX-ListQuestion can be found on [5]. LX-ListQuestion is available online at http://lxlistquestion.di.fc.ul.pt.

3 Comparing LX-ListQuestion and Other QA Systems

Comparing LX-ListQuestion with other QA systems is crucial to providing us with an assessment of how our system is positioned relative to the state-of-the-art. In this Section we compare the results of LX-ListQuestion with two other QA systems for Portuguese: RapPortagico and XisQuê.

The evaluation has two components: the quantitative evaluation of answers and the question coverage evaluation. The quantitative analysis uses precision, recall and F-measure as metrics. Nevertheless, these metrics do not accurately reflect how effective the systems are in providing correct answers to the maximum number of questions. For that, we use the question coverage, which determine the number of questions that receive at least one correct answer.

The question dataset used in these experiments is based on Páigico Competition[2]. The whole dataset is composed by 150 questions about Lusophony extracted from the Portuguese Wikipedia [4]. For the experiments, we use a subset of 30 questions whose expected answer type is Person or Location. We pick these two types since they are the ones more accurately assigned by the underlying a Named Entity Recognization tool named LX-NER [3]. Note, however, that our approach is not intrinsically limited to only these types.

3.1 Comparing Design Features

In this Section we compare the design features of LX-ListQuestion, RapPortagico and XisQuê. As mention at Sect. 2, LX-ListQuestion is a web-based QA system that finds a list of answers retrieving documents from the web and extracting candidates answers from inside the documents. RapPortagico [9], an off-line QA system that uses Wikipedia to retrieve the answers for List questions and XisQuê [1,2], a Web-based QA system that answers factoid questions that selects the most important paragraph of the Web pages and extracts the answer through the use of hand-built patterns. Table 1 shows the differences between the design features of the systems.

[2] http://www.linguateca.pt/Pagico/.

RapPortagico pre-indexes the documents using noun phrases that occur in the sentences in the corpus while LX-ListQuestion does not uses any pre-indexing of documents. RapPortagico uses the off-line Wikipedia as the source of information, while LX-ListQuestion uses the Web to find the answers. Both systems are also different in the type of answers. RapPortagico returns a List of Wikipedia pages and LX-ListQuestion returns a list of answers. The design features of LX-ListQuestion and XisQuê are to a certain extent similar. Both systems are Web-based QA systems and use the Web as the source of answers, and Google as supporting search engine. What differs between the systems is that the XisQuê answers Factoid Questions and LX-ListQuestion answers List Questions.

Table 1. Comparing Design Features of QA systems

	RapPortagico	XisQuê	LX-ListQuestion
Corpus pre-indexing	Yes. It pre-indexes the corpus using Noun Phrases	No	No
Corpus source	Off-line Wikipedia documents	Web	Web
Search engine	Lucene (indexed to documents stored into local files)	Google	Google
Type of questions	Factoid and List	Factoid	List
Type of answers	List of Wikipedia pages	Answer and Snippet	List of Answers

3.2 Quantitative Evaluation and Question Coverage

The evaluation was performed for the same set of questions for all systems. Table 2 shows the results of comparing the three systems. LX-ListQuestion obtained more correct answers than other systems, this can be verify in the recall measure with 0.120 and 0.097 and 0.055 respectively. However, it has lower precision since it returned more candidates than the other system. When comparing F-measure, LX-ListQuestion achieved slightly better results, obtaining 0.102 against 0.095 for RapPortagico and better results than XisQuê, that only obtained 0.078.

The question coverage that indicate the usefulness of the system to the user counting the number of questions for which the system provides at least one correct answer. Table 3 summarizes the number of questions answered by each system. From the 30 questions in the dataset, LX-ListQuestion provided at least one correct answer to 17 of them, against 14 of RapPortagico and 13 of XisQuê. The question coverage evaluation also allowed us to uncover an interesting behavior of these systems. For 7 questions answered by LX-ListQuestion, RapPortagico

Table 2. Evaluation of QA systems - LX-ListQuestion, RapPortagico and XisQuê

Experiments	Refer. answer list	Correct answers	All answers retrieved	Recall	Precision	F-Measure
LX-ListQuestion	340	41	460	0.120	0.089	0.102
RapPortagico		32	327	0.097	0.100	0.098
XisQuê		19	139	0.055	0.136	0.078

Table 3. Question Coverage

	LX-ListQuestion	RapPortagico	XisQuê
Number of Questions Answered	17	14	13

Table 4. Examples of answers provided by the each system

Question	Correct answers		
	LX-ListQuestion	RapPortagico	XisQuê
Cidades que fizeram parte do domínio português na India	Damao	Calecute	—
Cities that were part of the Portuguese Empire in India		Goa	—
Praias de Portugal boas para a práitica de Surf	Ericeira	—	Guincho
Good Portuguese beaches for surfing	Arrifana	—	Peniche
	Praia Vale Homens	—	
	São João Estoril	—	
Cidades Lusófonas conhecidas pelo seu Carnaval	Salvador	Mindelo Cabo Verde	Olinda
Lusophone cities known for their carnival celebrations	Recife	—	
	São Paulo	—	

Table 5. Results overview

Systems	Refer. answers list	Correct answers	All answers retrieved	Recall	Precision	F-Measure
LX-ListQuestion	340	41	460	0.120	0.089	0.102
RapPortagico		32	327	0.097	0.100	0.098
XisQuê		19	139	0.055	0.136	0.078
Combination		80	914	0.235	0.087	0.126

provided no answer. Conversely, for 5 questions answered by RapPortagico, LX-ListQuestion provided no answer. In addition, we note that when a question is answered by both systems, the answers given by each system tend to be different. Concerning XisQuê and LX-ListQuestion, we find that a large majority of correct answers given by XisQuê are different from those given by LX-ListQuestion. Namely, in 9 out of 13 questions to which XisQuê provides a correct answer, that answer is not present in the list of answers given by LX-ListQuestion. This result points towards a certain degree of complementarity between the systems. Table 4 shows some examples of questions and answers provided by each system that demonstrate the complementarity between the systems.

4 Concluding Remarks

In this paper we present an evaluation of our system, LX-ListQuestion, a Web-based QA system that uses redundancy and heuristics to answer List questions and compared the results with other two QA systems: RapPortagico and XisQuê. Our evaluation shows that our LX-ListQuestion achieved better results, with 0.102 in F-Measure, against 0.098 of RapPortagico, and 0.078 of XisQuê. The question coverage evaluation points towards a certain degree of complementarity between these systems. We observe that for a set of questions answered by LX-ListQuestion, the other systems provide no answers. Conversely, for some other questions answered by RapPortagico or XisQuê, LX-ListQuestion provided no answer. Based on our experiments, we noted that the approaches of RapPortagico, XisQuê and LX-ListQuestion may reinforce each other. To demonstrate these assumption, we built Table 5 with an overview of the results obtained in the experiments. The last row is the hypothetical combination of LX-ListQuestion, RapPortagico and XisQuê. As we can see, a QA system that combines their approaches can achieve better results and improve Recall and F-measure metrics.

References

1. Branco, A., Rodrigues, L., Silva, J., Silveira, S.: Real-time open-domain QA on the Portuguese web. In: Geffner, H., Prada, R., Machado Alexandre, I., David, N. (eds.) IBERAMIA 2008. LNCS (LNAI), vol. 5290, pp. 322–331. Springer, Heidelberg (2008)
2. Branco, A., Rodrigues, L., Silva, J., Silveira, S.: XisQuê: an online QA service for Portuguese. In: Teixeira, A., Lima, V.L.S., Oliveira, L.C., Quaresma, P. (eds.) PROPOR 2008. LNCS (LNAI), vol. 5190, pp. 232–235. Springer, Heidelberg (2008)
3. Ferreira, E., Balsa, J., Branco, A.: Combining rule-based and statistical models for named entity recognition of Portuguese. In: Proceedings of Workshop em Tecnologia da Informaçãoe de Linguagem Natural, pp. 1615–1624 (2007)
4. Freitas, C.: A lusofonia na Wikipédia em 150 topicos. Linguamatica **4**(1), 9–18 (2012)
5. Gonçalves, P.: Open-Domain Web-Based Multiple Document Question Answering forList Questions with Support for Temporal Restrictors. Ph.D. thesis, University of Lisbon, Lisbon, Portugal, 6 2015
6. Gonçalves, P., Branco, A.: Answering list questions using web as a corpus. In: Proceedings of the Demonstrations at the 14th Conference ofthe European Chapter of the Association for Computational Linguistics, pp. 81–84. Association for Computational Linguistics, Gothenburg, April 2014
7. Gonçalves, P., Branco, A.: Open-domain web-based list question answering with LX-listquestion. In: Proceedings of the 4th International Conference on WebIntelligence, Mining and Semantics, WIMS 2014, pp. 43:1–43:6. ACM, New York (2014)
8. Radev, D.R., Qi, H., Wu, H., Fan, W.: Evaluating web-based question answering systems. In: Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002. European Language Resources Association, Las Palmas, 29-31 May 2002

9. Rodrigues, R., Oliveira, H.: Uma abordagem ao páigico baseada no processamento e anáilise desintagmas dos tópicos. Linguamatica **4**(1), 31–39 (2012)
10. Strzalkowski, T., Harabagiu, S.: Advances in Open Domain Question Answering, 1st edn. Springer Publishing Company Incorporated, Netherlands (2007)
11. Voorhees, E.: Evaluating question answering system performance. In: Strzalkowski, T., Harabagiu, S. (eds.) Advances in OpenDomain Question Answering. Text, Speech and Language Technology, vol. 32, pp. 409–430. Springer, Netherlands (2006)