Information Processing and Management 54 (2018) 564–575

Contents lists available at ScienceDirect

Information Processing and Management

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

Do systems pass university entrance exams?

Alvaro Rodrigo*,[a], Anselmo Peñas[a], Yusuke Miyao[b], Noriko Kando[b]

[a] NLP & IR Group at UNED, Madrid, Spain
[b] National Institute of Informatics, Japan

ARTICLE INFO

Keywords:
Question answering
Reading comprehension
Semantic understanding

ABSTRACT

Reading Comprehension tests are commonly used to assess the degree to which people comprehend what they read. This is why we work with the hypothesis that it is reasonable to use these tests to assess the degree to which a machine "comprehends" what it is reading. In this work, we evaluate Question Answering systems using Reading Comprehension tests from exams to enter University. This article analyses the datasets generated, the kind of inferences required, the methodology followed in three evaluation campaigns, the approaches presented by participants and current results. Besides, we study the evolution of systems and the main lessons learned in this evaluation process. We also show how current technologies are unable to pass university-entrance exams. This is because these tests require a deep understanding of texts, as well as detecting the similar meaning of phrases with different words. Future directions focused on these ideas seem more promising than including a massive amount of data for training systems, what has allowed systems to obtain outstanding results in Reading Comprehension tests with more straightforward questions. We think this study helps to increase the knowledge about how to develop better Question Answering systems.

1. Introduction

Users continue demanding automatic systems able to extract precise information from the vast amount of digital documents currently available. Although there have been significant advances in the last decades, systems are still unable to answer complex questions. Besides, these systems usually return documents instead of short texts.

There have been several efforts focused on overcoming these issues. Some of these efforts have been based on proposing evaluation campaigns oriented to promote better systems. These campaigns have offered a framework where researchers can compare different approaches under the same conditions. Question Answering (QA) systems have been the focus of some of these campaigns. QA systems return exact answers to questions formulated in Natural Language.

The first QA evaluations, such as those at the Text REtrival Conference (TREC)[1], proposed to extract answers from large document collections (Voorhees & Tice, 1999). This format gave great importance to Information Retrieval in the context of QA systems and made it difficult to extract answers from texts with words different from those used in questions.

On the other hand, the Cross Language Evaluation Forum (CLEF)[2] proposed a QA task based on multiple-choice RC tests. The objective of this task was to promote systems able to detect semantic similarities among texts (Peñas et al., 2013). An important part of this proposal at CLEF was the definition of the Entrance Exams task, which is based on tests that humans have to pass to enter

* Corresponding author.
E-mail addresses: alvarory@lsi.uned.es (A. Rodrigo), anselmo@lsi.uned.es (A. Peñas), yusuke@nii.ac.jp (Y. Miyao), kando@nii.ac.jp (N. Kando).
[1] http://trec.nist.gov/.
[2] http://www.clef-initiative.eu/.

https://doi.org/10.1016/j.ipm.2018.03.002
Received 27 July 2017; Received in revised form 2 March 2018; Accepted 6 March 2018

University. This task can be seen as a Turing test where systems must emulate human performance (Turing, 1950). Entrance Exams proposed systems to answer questions from Reading Comprehension (RC) tests. This task includes some of the major current issues in QA, for instance, different rewordings between questions and documents (Momtazi & Klakow, 2015), complex questions (Chali, Hasan, & Mojahid, 2015), etc. These issues are also common in human tests.

Other researchers have also proposed RC tests for evaluating information-access systems (Hermann et al., 2015; Yang, Yih, & Meek, 2015). However, these researchers have not compared the performance of different systems under the same framework, as well as these researchers have paid little attention to the main drawbacks of current technologies.

For some of these tests (Rajpurkar, Zhang, Lopyrev, & Liang, 2016), the research community is making rapid progress with methods based on deep learning and extensive data (Hu, Peng, & Qiu, 2017). These methods achieve better results than those reported in this work[3]. Nevertheless, these tests are different from those used in this paper. Our tests are more complex because they contain candidates very close to correct answers and require systems to apply several inferences from texts, which makes it difficult to find the right answer. The analysis of our questions shows that systems need more than processing massive data to find correct answers to such complex questions.

1.1. Research objective

The primary objective of this article is to study the results of information-access systems doing a human task: university-entrance examinations. More in detail, we want to analyze the performance of current technologies over real data designed for humans. Then, we want to know the main difficulties for current approaches to propose new directions for improving systems' results.

In summary, our study aims to increase research over QA systems from the perspective of evaluation, addressing the following four research questions:

- **RQ1:** How can we adapt human tests for evaluating information-access systems?
- **RQ2:** How well do current systems perform over those human tests?
- **RQ3:** What are the main issues for current technologies when answering those tests?
- **RQ4:** What should be included in new evaluations aimed to promote better systems?

Despite the difficulty of the challenge, the proposed task set a benchmark for evaluating technologies aimed to answer questions about a document, finding semantic similarities among texts, etc. In this paper, we describe the task and analyze collections and results to find the main drawbacks for current technologies. The paper is organized as follows: we introduce the main Related Work in Section 2. Section 3 describes the methodology of our work, while Section 4 describes and analyzes the data employed. Then, we describe in Section 5 how systems are evaluated. The analysis of results is given in Section 6. The main approaches applied by participant systems are reported in Section 7. In Section 8, we describe the main lessons learned in this work. Finally, we give some conclusions and future work in Section 9.

2. Related work

In this Section, we describe similar efforts for evaluating QA technologies. We start with a short description of the first evaluations over free-text, and then we move to new proposals based on RC.

Evaluations at TREC represented an essential event in QA research (Voorhees & Tice, 1999). Similar campaigns were proposed at CLEF and the NII Test Collection for Information Retrieval (IR) Systems (NTCIR)[4], including multilingual and cross-lingual tasks (Fukumoto, Kato, & Masui, 2004; Magnini et al., 2004). These campaigns were based, as IR evaluations, on the Cranfield paradigm (Cleverdon, 1967). They prepare a document collection and a set of factoid questions, for example "*Who is the president of X?*" or "*When was Y born?*". Then, participant systems returned answers extracted from documents. However, these evaluations focused on *acknowledging only correct answers* without penalizing incorrect ones, which promoted systems based on IR which returned answers no matter whether those answers were incorrect.

With the purpose of leaving aside the strong dependency on IR components and improving results, CLEF decided to move the traditional QA task to a new one based on RC tests. For this purpose, CLEF proposed the Question Answering for Machine Reading Evaluation (QA4MRE) task, whose first edition was held in 2011 (Peñas et al., 2012). The idea behind this change was that if RC tests are suitable for assessing document understanding, they can be suitable for evaluating QA systems. QA4MRE employed RC tests created only for the task. However, these tests were a little bit artificial and complicated, and systems' results were very poor. These observations motivated a change towards a new task named Entrance Exams, which is described in this paper (the main details about methodology are in Section 3).

On the other hand, there have been other proposals based on assessing document understanding out of the scope of evaluation campaigns. Some researchers studied if RC tests were suitable for evaluating machine "intelligence" (Clark & Etzioni, 2016). These researchers proposed different levels of inference that can be tested using RC collections. Besides, these researchers recommended creating tests for machines, given that the objectives of human tests are different. We think it is valuable to evaluate systems also

[3] A board with results for the SQuAD dataset is in https://rajpurkar.github.io/SQuAD-explorer/.
[4] http://research.nii.ac.jp/ntcir.