Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 275 – 284

Procedia Computer Science

AQA-WebCorp: Web-based Factual Questions for Arabic

Wided BAKARI[a*], Patrice BELLOT[b] Mahmoud NEJI[c]

[a]*Faculty of Economics and Management, 3018, Sfax Tunisia, MIR@CL, Sfax, Tunisia*
[b]*Aix-Marseille University, F-13397, Marseille Cedex 20, LSIS, Marseille, France*
[c]*Faculty of Economics and Management, 3018, Sfax Tunisia, MIR@CL, Sfax, Tunisia*

Abstract

Working with corpus construction becomes an interesting alternative to different applications of natural language processing, such as, question-answering, machine translation, information retrieval, etc. Similarly, with the heterogeneous data and the user demands for the accurate information, many studies have accentuated the need of the Web to highlight the corpus construction. As well as, Arabic doesn't have an equivalent number of linguistic corpuses as compared to other languages like English. In this paper, we focus on building our corpus of Arab questions-texts. We present a method for recovering text passages. This method is based on a real automatic interrogation of Google, in order to generate passages of texts and answer the factual questions. The first part of this paper describes the formal details about this method; the second part presents some experiments and results that validate our method.

Peer-review under responsibility of KES International

Keywords: Arabic, Corpus, Question analysis, Passage, Google, Corpus construction.

1. Introduction

A corpus is a collection of pieces of texts in electronic forms, selected according to external criteria for end to represent, if possible, a language as a data source for linguistic research [1]. Indeed, a definition that is both specific and generic of a corpus according to [2] is the result of choices that brings the linguists. A corpus is not a simple object; it should not be a mere collection of phrases or a "bag of words". This is in fact a text assembly that can cover many types of text. The construction of a corpus is not an easy task; it is a task that both essential and delicate. Also, it is complex because it depends in large part a significant number of resources to be exploited. One way to diminish this problem is using the Web as data sources. Indeed, the Web is a colossal quantity of texts was recovered freely [5]. It contains billions of text words that can be used for any kind of linguistic research [6]. Additionally, with the internet development and its services, the web has become a great source of documents in different languages and different areas. This source is combined with storage media that allow the rapid construction

* Corresponding author. Tel.: +21696286145;
E-mail address: wided.bakkari@fsegs.rnu.tn

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.140

of a corpus [3]. In addition, using the Web as a base for the establishment of textual data is a very recent task. The recent years have taken off work attempting to exploit this type of data. From the perspective of automated translation in [4], the others study the possibility of using the websites which offering information in multiple languages to build a bilingual parallel corpus.

Consequently, with the development of electronic media and the heterogeneity of Arabic data on the Web, the idea of building a clean corpus for certain applications of natural language processing, including machine translation, information retrieval, question answer, become more and more pressing. Arabic is also an international language, rivaling English in number of native speakers. However, little attentions have been devoted to this language. Although there have been a number of investigations and efforts invested the Arabic corpus construction, especially in Europe; progress in this area is still limited. In fact, there are few publicly available corpuses, especially for Arabic. The lack and / or the absence of corpus in Arabic have been a problem for the implementation of natural language processing. This also has a special interest in the track of the question answering. Today, the Web has been a driving force in innovations within information retrieval, as users worldwide use search engines to find relevant content on the web. For question-answering, information retrieval methods are used for retrieving documents relevant to the question, and selecting documents likely containing the answer. Most question-answering systems use existing search engines.

In our research, we completed building our corpus of questions-texts AQA-WebCorp (Arabic Question-Answering Web Corpus) by querying the search engine Google. Google has been working on several initiatives to help increase Arabic-language content. Notably, we are concerned our aim, a kind of, giving a question, analyzing texts at the end to answer this question. Therefore, we seek to create and develop our own corpus of pair's questions-texts. This constitution then will provide a better base for our experimentation step. Thus, we try to model this constitution by a method for Arabic insofar as it recovers texts from the web that could prove to be answers to our factual questions. So that, this paper is organized into six sections as follows: it begins with an introduction, followed by the challenges of the Arabic language. Section 3 outlines the earlier work in Arabic; Section 4 shows our proposed method to build a corpus of pairs of questions and texts; Section 5 describes an experimental study of our method; a conclusion and future work will conclude this article.

2. The challenges of the Arabic language

Although, Arabic is within the top ten languages in the internet, it lacks many tools and resources. Meanwhile, Arabic language is the official language in all Arab nations as Tunisia, Egypt, Saudi Arabia and Algeria. Moreover, it is also an official language in non-Arab countries as Chad and Eritrea. The Arabic does not have capital letters compared the most Latin languages. This issue makes so difficult the natural language processing, such as, named entity recognition. Unfortunately there is very little attention given to Arabic corpora, lexicons, and machine-readable dictionaries [20]. In their work [21], the authors suggest that the developed Arabic question-answering systems are still few compared to those developed for English or French, for instance. This is mainly due to two reasons: lack of accessibility to linguistic resources and tools, such as corpora and basic Arabic NLP tools, and the very complex nature of the language itself (for instance, Arabic is inflectional and non concatenative and there is no capitalization as in the case of English). On their part, [22] illustrate some difficulties of Arabic. This language is highly inflectional and derivational, which makes its morphological analysis a complex task. Derivational: where all the Arabic words have a three or four characters root verbs. Inflectional: where each word consists of a root and zero or more affixes (prefix, infix, suffix). Arabic is characterized by diacritical marks (short vowels), the same word with different diacritics can express different meanings. Diacritics are usually omitted which causes ambiguity. Absence of capital letters in Arabic is an obstacle against accurate named entities recognition. And then, in their survey [23], the authors emphasize that as any other language, Arabic natural language processing needs language resources, such as lexicons, corpora, treebanks, and ontologies are essential for syntactic and semantic tasks either to be used with machine learning or for lookup and validation of processed words.