Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 130 (2018) 589–596

Procedia
Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Ambient Systems, Networks and Technologies, ANT-2018 and the 8th International Conference on Sustainable Energy Information Technology, SEIT 2018, 8-11 May, 2018, Porto, Portugal

Semantic Annotation of Arabic Web Documents using Deep Learning

Saeed Albukhitan, Ahmed Alnazer, Tarek Helmy [a,*]

Information and Computer Science Department,
College of Computer Science & Engineering,
King Fahd University of Petroleum & Minerals,
Dhahran 31216, Mail Box 413, Saudi Arabia,
[a] On leave from College of Engineering, Department of Computers & Automatic Control,
Tanta University, Egypt
[albokhitan,alnazera,helmy]@kfupm.edu.sa

Abstract

The vision of Semantic Web is to have a Web of things instead of Web of documents in a form that can be processed by machines. This vision could be achieved on the existing Web using semantic annotation based on common and public ontologies. Due to exponential growth and the huge size of the Web sources, there is a need to have fast and automatic semantic annotation of Web documents. Arabic language received less attention in semantic Web research as compared to Latin languages especially in the field of semantic annotation. The aim of this paper is to investigate the feasibility of using word embeddings from deep learning algorithms for semantic annotation of Arabic Web documents. To evaluate the performance of the proposed framework, food, nutrition, and health ontologies were used to annotate some related Web documents._For a given set of Arabic documents and ontologies, the framework produces annotations of these documents using different output formats. The initial results show a promising performance which will support the research in the Semantic Web with respect to Arabic language. The proposed framework could be used for building semantic Web application and semantic search engines for Arabic Language.

Keywords: Deep Learning; Semantic Annotation; Arabic Language; Ontology

* Corresponding author. Tel.: +966 -13-8601967; fax: +966-13-8602174.
E-mail address: helmy@kfupm.edu.sa

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2018.04.108

1. Introduction

The Semantic Web extends the current Web content with meta-data that can be understood by machines[1]. To achieve successfully the goal of semantic Web, it is required to have the adequate quantity of related semantic with content in high-quality. Semantic annotation of the existing Web is one way to generate such content. Semantic annotation of Web sources is a process that adds machine-understandable content to the natural language textual data. Semantic annotation of Web contents manually is not feasible or scalable due to the huge amount and the rate of emerging Web content.

For semantic Web application to make use of a Web data source, the data needs to be transformed into a machine understandable format[2]. Semantic Web technologies provide different means to represent the information for machine processing. The most common format is RDF for the extracted information and Web Ontology Language (OWL) for ontological representation of concepts, their relationship and semantic rules that could be applied to the knowledge.

Most of Web semantic annotation tools only support Latin languages and very hard to adapt to Arabic language. The amount of research and work done for annotating Arabic content in Web is very limited and non-scalable.. One of the biggest challenges facing Arabic research is the availability and accessibility of Arabic resources, such as ontologies, corpora, named list, dictionaries, and NLP tools. This shortage makes the collection, analyzes, and investigation of such resources laborious especially if the semantic annotation techniques depend on such resources.

Deep Learning (DL) has proved to provide a good improvement in multiple areas including text mining[16]. By using DL, it is possible to have word embedding as distributed word representations from textual data by applying neural language models like CBOW and Skip-gram. The application of DL to aid semantic annotation of Arabic documents remains largely unexplored.

The identification of named entities in the domain text has proven difficult due to term variation and term ambiguity because a concept can be expressed by various realizations (a.k.a. term variants). To resolve this issue, a large-scale database that contains longer words and phrases as well as shorter forms like abbreviations or acronyms must be used. Finding all the term variants in the text is important for improving the results of Information Retrieval (IR) and Information Extraction (IE) systems, which traditionally rely on keyword-based approaches. Therefore, the number of documents retrieved is prone to change when using acronyms instead of and/or in combination with full terms.

This paper investigates the feasibility of using deep learning, an emerging area of artificial neural networks, for identifying ontological concepts in Arabic text. More specifically, we propose to use the two neural language models Skip-gram and CBOW (Continuous Bag-of-Words) to produce word embeddings, which are distributed word representations typically induced using neural language models. These word embeddings can be linked to classes formalized in domain Ontology represented in the W3C Web Ontology Language (OWL). We present complete semantic annotation framework based on deep learning for Arabic Web documents that annotates the Arabic Web resources and produces annotation. It supports parsing of ontologies stored in a different format including RDF, OWL, and N-TRIPLE.

The rest of this paper is organized as follows. Section 2 reviews the existing related work on semantic annotation and deep learning. Section 3 presents the proposed framework. Section 4 presents experimental setup and evaluation with the discussion in Section 5. Finally, we summarize the paper and highlight the future work directions in Section 6.

2. Related Work

A survey of some semantic Web technologies supporting Arabic could be found in Beseiso et al.[1]. Four mostly used semantic Web tools were investigated, namely Protégé, Jena, Sesame, and KOAN. Their investigation focuses on the tool's functionality, type of standards supported and support level of Arabic language. Moreover, those tools do not support Arabic language completely as compared to Latin languages. The most supporting tool for Arabic language was Jena with some limited support for query processing. Arabic language is complex comparing to Latin languages. The common challenges of Arabic language with respect to NLP tasks were highlighted in Abdel Rahman et al.[2]. Arabic language does not have features such as case-sensitivity which is an important feature used by Latin languages to detect proper names. Arabic words could have more than one affix and can be expressed as a combination