Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 127 (2018) 82–91

Procedia Computer Science

www.elsevier.com/locate/procedia

The First International Conference On Intelligent Computing in Data Sciences

Data science in light of natural language processing: An overview

Imad Zeroual[a]* and Abdelhak Lakhouaja[a]

[a]*Faculty of Sciences, Mohamed First University, Av Med VI BP 717, Oujda 60000, Morocco*

Abstract

The focus of data scientists is essentially divided into three areas: collecting data, analyzing data, and inferring information from data. Each one of these tasks requires special personnel, takes time, and costs money. Yet, the next and the fastidious step is how to turn data into products. Therefore, this field grabs the attention of many research groups in academia as well as industry. In the last decades, data-driven approaches came into existence and gained more popularity because they require much less human effort. Natural Language Processing (NLP) is strongly among the fields influenced by data. The growth of data is behind the performance improvement of most NLP applications such as machine translation and automatic speech recognition. Consequently, many NLP applications are frequently moving from rule-based systems and knowledge-based methods to data-driven approaches. However, collected data that are based on undefined design criteria or on technically unsuitable forms will be useless. Also, they will be neglected if the size is not enough to perform the required analysis and to infer the accurate information. The chief purpose of this overview is to shed some lights on the vital role of data in various fields and give a better understanding of data in light of NLP. Expressly, it describes what happen to data during its life-cycle: building, processing, analyzing, and exploring phases.

Keywords: Data science; Natural language processing; Data driven approches; Corpora; Machine learning

1. Introduction

Recently, a variety of Natural Language Processing (NLP) applications are based on data-driven methods such as neural networks and Hidden Markow Models (HMMs) [1]. Since the progress in most of these methods is driven

* Corresponding author. Tel.: +212618417420.
E-mail address: mr.imadine@gmail.com

10.1016/j.procs.2018.01.101

from data, large and high-quality data become very valuable resources. For instance, some of their beneficial effects have been observed in machine translation [2], word sense disambiguation [3], summarization [4], syntactic annotation [5], named entity recognition [6], among other NLP applications. Over centuries, the primary data have accumulated manually in the form of unstructured repositories of texts called archives. In the early 1980s, the revolution of computer industry leads to a new way of data treatment primarily in term of storage capacity. The new form of an entire population of electronic data are called databases that are designed to facilitate data entry and retrieval. A few years later, the NLP scientists gave a name of corpora to a subset of databases. i.e., a collection of naturally occurring text samples are compiled according to some explicit criteria in order to represent a language [7].

Similarly, various fields such as linguistic, lexicography, and education have been supported by data. This latter, in form of archives, dictionaries, and corpora, are considered as a principal source of evidence for linguistic description and argumentation. Grammarians have always needed sources of evidence as a basis to illustrate grammatical features such as the nature, the structure, and the functions of languages [8]. In lexicography, corpora are exceedingly used to build and make major revision of relevant dictionaries such as the "Dictionary of the Older Scottish Tongue", the "Middle English Dictionary", the "Dictionary of Old English", and the "Oxford English Dictionary" [9]. Alongside the linguistic description and lexicography, data significantly affect a wide range of research activities that have a pedagogical purpose. For instance, learner corpora, collections of first or second language learner data, are generally used to build word frequency lists. These lists are a quick guide and better curricula materials for teaching and learning vocabulary. *Nation* [10] believes that the high-frequency words is important for the learners and need to focus on their learning burden and ensure that the learners will come back to them again. Whereas, some low frequent words may not need to become a part of the learners' output or the teacher may give some brief attention to them.

The aim of this paper is to outline the main stages in life-cycle of data in NLP, from data design and collection to data processing and analysis. Further, to make this overview equally rich in both theoretical and practical aspects, a survey[a], that covers 100 of well-known and influential corpora, has addressed these stages. Yet, it represents many languages including monolingual corpora (24 languages), bilingual corpora (11 languages), and multilingual corpora (3 to 109 languages). Since the English language was the forerunner in NLP, it is normal that 25% of covered corpora are devoted to English. However, many other languages are catching up, implicitly considering English corpora as a global standard. It worth mention that most covered corpora are publicly available, either free of charge or at an affordable cost, and some of them are available for online search or downloadable. Finally, information regarding the website addresses or *DOI* for all data mentioned in this survey are given in the appendix.

After this introduction, the main content of the paper is structured in section 2. We outline various stages of data life-cycle in light of NLP. Expressly, data sources, format, and corpus design criteria are described. Further, we attempt to present a general view of different data collection methods. Then, data processing and analysis are discussed in detail. Finally, Section 3 contains some concluding remarks.

2. Life-cycle of data in NLP

NLP, also known as computational linguistics, is a subfield of artificial intelligence that aims to learn, understand, recognize, and produce human language content [11]. To get over the limitation of rule-based systems, many research groups move to data-driven methods. However, collecting data is not a trivial undertaking and demands a coherent treatment behind it. For instance, one of the biggest challenges all corpora builders encounter is the lack of public resources as well as copyright. Yet, the quality and quantity of data does matter and should not haphazardly collected. For instance, *Manning* [12] reports that what we need to enhance the part-of-speech tagging and move the accuracy from its current level of about 97.3% to close to 100% is using better training data.

2.1. Data design, sources, and format

A corpus is not haphazard collections of textual material, therefore, a great care must be taken during the data collection [13]. Building corpora usually starts by identifying the appropriate criteria in order to be representative

[a] https://goo.gl/forms/PwxGLe7imVChBPEx2