Pattern Recognition Letters 41 (2014) 93–102

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A graph-based multi-level linguistic representation for document understanding ☆

David Pinto [a,*], Helena Gómez-Adorno [a,1], Darnes Vilariño [a,1], Vivek Kumar Singh [b,2]

[a] Benemérita Universidad Autonóma de Puebla, Faculty of Computer Science, 14 Sur & Av. San Claudio, CU, Edif. 104C, Puebla, Mexico
[b] South Asian University, Department of Computer Science, Akbar Bhawan, Chanakyapuri, New Delhi 110021, India

ARTICLE INFO

Article history:
Available online 9 December 2013

Keywords:
Text mining
Text representation
Graph-based representation

ABSTRACT

Document understanding goal requires discovery of meaningful patterns in text, which in turn requires analyzing documents and extracting information useful for a purpose. The documents to be analyzed are expected to be represented in some way. It is true that different representations of the same piece of text might have different information extraction outcomes. Therefore, it is very important to propose a reliable text representation schema that may incorporate as many features as possible, and at the same time provides use of efficient document understanding algorithms. In this paper, we propose a graph-based representation of textual documents that employs different levels of formal representation of natural language. This schema takes into account different linguistic levels, such as lexical, morphological, syntactical and semantics. The representation schema proposed is accompanied with a proposal for a technique which allows to extract useful text patterns based on the idea of minimum paths in the graph. The efficiency of the representation schema proposed has been tested in one case of study (Question-Answering for machine Reading Evaluation – QA4MRE), and the results of experiments carried in it, are described. The results obtained show that the proposed graph-based multi-level linguistic representation schema may be successfully used in the broader framework of document understanding.

1. Introduction

A huge amount of information produced on a daily basis is found in different forms of natural language written texts, such as magazines, books, e-books, journals, technical reports, etc. In fact, we are now overwhelmed with textual data, which increases every other day. The explosive growth in the number of such documents needs development of effective approaches to explore, analyze, and discover knowledge from documents. Developing automated tools for machine reading by discovering patterns and extracting knowledge from texts is one of the most important goals of Text Mining (TM) research. And the usual assumption in it is that texts are represented in some kind of structure.

The present research work is mainly concerned with the construction of a suitable text representation model based on graphs, that can facilitate discovering of important text patterns from it. We propose to state and demonstrate that the features (text patterns) so discovered can be used in different tasks associated to document understanding (such as for document classification, information retrieval, information filtering, information extraction and question answering).

The text pattern discovering technique proposed here is based on the traversal of the graph representation of documents, using the shortest paths. This text pattern discovery is used in our experimental case study for estimating similarities between pairs of texts. The case study of question answering validation for reading comprehension tests presented here demonstrates the working and efficacy of our framework. The results of experimental work reported are analyzed and key observations clearly stated.

In summary, this research work presents a new text representation schema useful for mining documents, exploiting their lexical, syntactic, morphologic and semantic information. The representation schema is built over a syntactic analysis developed through a dependency parser for all the sentences in the document, including further morphologic and semantic information. The final result obtained is an enriched output in the form of a graph that represents the input document in the form of a multiple level formal representation of natural language sentences. The graph-based representation schema and the similarity measure proposed here, enables a more effective and efficient text mining process.

The rest of the paper is organized as follows. Section 2 presents a literature survey on the different text representation schemata

☆ This paper has been recommended for acceptance by J. Fco. Martínez-Trinidad.

* Corresponding author. Tel.: +52 222 2295500x2856.
E-mail addresses: dpinto@cs.buap.mx (D. Pinto), helena.adorno@gmail.com (H. Gómez-Adorno), darnes@cs.buap.mx (D. Vilariño), vivekks12@gmail.com (V.K. Singh).
[1] Tel.: +52 222 2295500x2856.
[2] Tel.: +91 11 24195148.

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.12.004

proposed. It also emphasizes the contribution of using graph-based structures in the text representation research field. Section 3 explains in detail the graph-based text representation schema proposed. The diverse features that may be included into this representation are discussed along with suitable examples. Section 4 describes our proposal of an efficient method for discovering texts patterns from the graph-based representation of text documents. Section 5 presents the performance assessment of the proposed schema of text representation, in the particular case study of QA4MRE. It first describes the task and then illustrates the process of discovering text patterns. Finally, the results obtained in the experiments are reported. Section 6 concludes the paper by presenting the main contribution and findings of this research work.

2. State of the art

The most conventional text representation schemata observed in applications like information retrieval, text categorization, authorship attribution etc. are: Bag of Words (BoW) (Mladenic and Grobelnik, 1998), *n*-grams model (Stamatatos et al., 2001; Keselj et al., 2003), boolean models (Mauldin, 1991), probabilistic models (Croft et al., 1991) and vector-space models (Salton, 1988). The majority of these text representations are based on the BoW representation, thus ignoring the words sequentiality and, hence, the meaning implied or expressed in the documents as well. This deficiency generally results in failure to perceive contextual similarity of text passages. This may be due to the variation of words that the passages contain. Another possibility is perceiving contextually dissimilar text passages as being similar, because of the resemblance of their words.

For many problems in natural language processing, a graph structure is an intuitive, natural and direct way to represent the data. There exist several research works that have employed graphs for representing text. A comprehensive study of the use of graph-based algorithms for natural language processing and information retrieval can be found in Mihalcea and Radev (2011). It describes approaches and algorithmic formulations for: (a) synonym detection and automatic construction of semantic classes using measures of graph connectivity on graphs built from either raw text or user-contributed resources; (b) measures of semantic distance on semantic networks, including simple path-length algorithms and more complex random-walk methods; (c) textual entailment using graph-matching algorithms on syntactic or semantic graphs; (d) word-sense disambiguation and name disambiguation, including random-walk algorithms and semi-supervised methods using label propagation on graphs; and (e) sentiment classification using semi-supervised graph-based learning or prior subjectivity detection with min-cut/max-flow algorithms. Although the work described in Mihalcea and Radev (2011) covers a wide number of algorithms and applications, there exist other relevant works in literature worth mentioning. A great interest has grown in the computational linguistic community for using this kind of text representation in diverse tasks of natural language processing, such as in summarization (Zha, 2002), coreference resolution (Nicolae and Nicolae, 2006), word sense disambiguation (Dorow and Widdows, 2003; Veronis, 2004; Agirre et al., 2006), word clustering (Matsuo et al., 2006; Biemann, 2006), document clustering (Zhong, 2005), etc. The majority of the approaches presented in literature use well known graph-based techniques in order to find and exploit the structural properties of the graph underlying a particular dataset. Because the graph is analyzed as a whole, these techniques have the remarkable property of being able to find globally optimal solutions, given the relations between entities. For instance, graph-based methods are particularly suited for disambiguating word sequences, and they manage to exploit the interrelations among the senses in the given context. Unfortunately, most of the research works that use graph-based representations propose ad hoc graph-structures that only work with the particular problem they are dealing with. It is, therefore, imperative to attempt to propose a general framework that may be used in different contexts with a minimum amount of changes.

3. A graph-based multi-level linguistic representation schema for documents

This section presents our proposed text representation schema that utilizes multiple linguistic levels of formal definition of natural language texts. The motivation for the schema is to capture most of the features present in a document, ranging from lexical to semantic level. By including lexical, syntactic, morphologic and semantic analysis in the representation, we attempt to represent how different text components (words, phrases, clauses, sentences, etc.) are related.

A labeled di-graph denoted by $G = \{V, E, L_V, L_E, \alpha, \beta\}$ is the starting point for representing the different levels of language description. Here:

- $V = \{v_i | i = 1, \ldots, n\}$ is a finite set of vertices, $V \neq \emptyset$, and n is the number of vertices in the graph.
- $E = \{(v_i, v_j) | v_i, v_j \in V, 1 \leqslant i,j \leqslant n\}$. Note that the notation (v_i, v_j) indicates that a given order is established.
- L_V is the tag set for the vertices.
- L_E is the tag set for the edges.
- $\alpha : V \rightarrow L_V$ is a function that assigns tags to vertices.
- $\beta : E \rightarrow L_E$ is a function that assigns tags to the directed edges.

The representation of each linguistic level together with their association with the graph components is described as follows.

3.1. Lexical level

At the lexical level we deal with words, one of the most basic units of text, describing their meaning in relation to the physical world or to abstract concepts, without reference to any sentence in which they may occur. Lexical definition attempts to capture everything that a term is used to refer to and, as such, is often too vague for many purposes. Therefore, it is used as a basic representation which need to be further enriched through higher levels of language description.

To illustrate the lexical level of representation, let us consider the following example sentence:

Text mining searches patterns in texts.

Thus, given a di-graph $G = \{V, E, L_V, L_E, \alpha, \beta\}$, the function α assigns lexical words to the vertices. In this case, the L_V set (set of all the lexical words found in the document to be represented) is $L_V = \{$"*Text*", "*mining*", "*searches*", "*patterns*", "*in*", "*texts*"$\}$. At this point, we have only assigned lexical components to the vertices of the graph, thus, the edges are not defined yet. In other words, there are no edges to reflect any relationship among the words in the graph. This is a basic representation that it is barely useful for practical purposes. Therefore, we move ahead to capture and represent the morphological level details of the language description.

3.2. Morphological level

At the morphological level we deal with the identification, analysis and description of the structure of a given language's morphemes and other linguistic units, such as root words, affixes and Parts of Speech (PoS). In order to introduce these morphological