Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 93 (2016) 521 – 532

Procedia
Computer Science

6th International Conference On Advances In Computing & Communications, ICACC 2016, 6-8 September 2016, Cochin, India

Evaluation and Analysis of Grammatical Linguistic Pattern over Social Science and Technology Textbooks

Phub Namgay[a], Anu Singha[b,]*

[a]*Department of Information Technology, Royal Thimphu College, Thimphu-00975, Bhutan*
[b]*Department of Computer Science, South Asian University, New Delhi, Delhi-110021, India*

Abstract

Every textbook is built upon the foundation of key concepts. Books that contain concepts that share some common properties and are semantically related are more lucid and intelligible than those that contain many unrelated concepts. These key concepts follow a certain widely used grammatical linguistic patterns. An enormous amount of information can be derived regarding these key concepts such as their dispersion across chapters and, the relationship among the concepts, etc. The relationship among key concepts can be used to evaluate the books, and draw concept graphs. Since we live in an increasingly visual society, graphic representation of the key concepts may well help readers in easily understanding textbook content.This paper describes an experiments performed on selected chapters from social science and technology textbooks. The goal of the experiments was to evaluate the effectiveness of the grammatical linguistic pattern in identifying the key concepts in these textbooks and to quantify the suitability of linguistic patterns in different fields.

Peer-review under responsibility of the Organizing Committee of ICACC 2016

*Keywords:*Key concepts; Linguistic pattern; Stanford POS tagger; Microsoft Web N-gram services; Wikipedia dump; Concept graph

* *AnuSingha*. Tel.: +*91*-8974310941.
E-mail address:anusingh5012@gmail.com

Peer-review under responsibility of the Organizing Committee of ICACC 2016
doi:10.1016/j.procs.2016.07.247

1. Introduction

With an increasing amount of textual data available, grammatical linguistic patterns have proven their importance in the text mining and knowledge engineering field[1]. There are many linguistic patterns widely used in natural language processing (NLP) and each grammar follows a certain pattern. Linguistic patterns are the ways in which various parts of speech are associated with each other.

The goal of the paper is to evaluate and analyze the grammatical linguistic patterns in social science and technology textbooks based on the key concepts used therein. The key concepts in our context correspond to the terminological noun phrases (e.g., the word *cognitive dissonance* is an adjective followed by noun. So the grammatical pattern is AN, where A is adjective and N is noun).This research is based on the intuition that a book that contains the right set of related key concepts is more comprehensible for the reader than those that do not. The set of key concepts in a textbook can further be used to draw concept graphs. We implement text mining techniques proposed in paper[2] and assess their effectiveness in analyzing textbooks in two different subject's viz., social science and technology. The choice is governed by the intuition that the metrics for quantifying a social science book will be different from those for quantifying a technology book. This difference can be attributed to the differences in the usage of words in defining the key concepts in the two domains. Our overall approach adopted in tackling this problem consists of the following steps: (1) Identifying the key concepts in chapters of the textbooks using a linguistic pattern (2) Pruning of the malformed and common knowledge key concepts (3) Analyzing the key concepts to draw useful conclusion and (4) Drawing a concept graph of the key concepts.

2. Related work

Much work has been done on grammatical linguistic patterns in the field of NLP. Algorithms on grammatical linguistic patterns are based on key concepts in a textbook. The key concepts are extracted using linguistic patterns. In this paper, we used the linguistic pattern P_3proposed in paper[2]:

$$P_3 = A^* N^+ \tag{1}$$

A and N represent Adjectives and Nouns in the sentence respectively. * represents zero or more terms. + represents one or more terms. Presently, there are two popular linguistic patterns widely used in the NLP[2]:

$$P_1 = C^* N \tag{2}$$

and

$$P_2 = (C * NP)^? (C * N) \tag{3}$$

From a comparison of linguistic patterns carried out in paper[3], the authors concluded that pattern P_1 outperform pattern P_2 in maximal pattern matches. They also observed that pattern P_3 performs better than pattern P_1 although only slightly. The impressive performances of pattern P_3 continues to hold after Microsoft Web N-gram Services pruning. With this successful experimental result and the good performance of grammatical linguistic pattern P_3 in determining the key concepts, we decided to adopt linguistic pattern P_3 in this paper. With pattern P_3, we are typically interested in phrases containing noun, adjectives and sometimes prepositions.

3. System overview

Fig. 1 demonstrates the sequential step by step implementation of all the tasks carried out in this research. The first task is to tag each word in the text file using a part of speech (POS) tagger and to evaluate the tagged file. This evaluation is important as all the subsequent text processing is based on the input from tagging. The linguistic pattern adopted is used to extract the key concepts from the tagged text file. The key concept extraction is carried out by the JavaCC parser. To eliminate disambiguates in tagging, the WordNet lexical database is used to correct tagging errors. With the above processing steps, the set of noun phrases obtained is likely to contain malformed and common knowledge phrases. So, further Microsoft Web N-gram Services pruning is carried out to extract a good set of noun phrases. With this final set of key concepts, detailed experimental analyses are carried out to produce useful information.