Intelligent Synoptic Discovery and Semantic Annotation of Ubiquitous Library Literature

Sabar, M.I.

Dept. Software Engineering, Faculty of Computing and Technology,
University of Kelaniya,
ishans@kln.ac.lk

Abstract

Key to all academic ventures and scholastic exploits is bibliographic reference. Traditionally however, the eager student has been burdened by the significant toll involved in researching for referential content. Volumes not available in the local physical libraries, have to be searched for online. Text-related indexes are abstract and terse, often not fully-reflective of the actual backend content, its depth, breadth, and/or knowledge-rending quality. Critical comparison of multiple texts of the same subject area are near-impossible. This research presents a novel, intelligent technique for semantic evaluation and synoptic discovery in ubiquitous library texts, both physical and online, affording efficacious volume searching, comparison, and optimal selection. This paper is organised into multiple sections. The first section provides an introduction to this research, the second section enunciates the problem statement, the third section states the objectives of this research, the fourth section provides the methodology used to achieve the findings herein, the fifth section asserts the significance of this research, the sixth section provides the results and the discussion, the seventh section asserts the conclusions and recommendations, and the eighth section lists the key references used.

Keywords: *intelligent, referential, metadata, synchronous, synoptic*

Introduction

Students typically devote long hours in physical libraries or online, perusing volumes of reference literature looking for the right text. The showcased reference volumes often do not provide adequate outlines of

their inherent content, depth, breadth, or knowledge value; the respective indexes simply paraphrasing contained topics. Clearly, this traditional researching technique needs revamping; intelligent synoptic discoveries in reference texts is key to facilitating optimal volume selection.

Problem Statement

How can an intelligent, online, library catalog be enacted which would synchronously generate synoptic annotations of the earmarked texts, enabling real-time, critical analysis and optimal text selection?

Objectives

The objectives of this research are:

- Enable critical comparison and evaluation of candidate texts.
- Enable machine-driven referrals of suitable literature.
- Afford foolproof, precise, machine-endorsed optimal text selection.

Materials and Methods

The Title, Index, and Introduction sections of each candidate text is parsed in turn, and preprocessed as follows.

Syntactic Recasting

Discard all occurrences of *Conjunctions* (2) in the parsed text.

Coordinating Conjunctions - eg., *for, and, nor, but, or, yet,* and *so*.

Correlative Conjunctions - eg., *both/and, either/or, neither/nor,* etc.

Subordinating Conjunctions - eg., after, although, as, as if, as long as, as much as, etc.

Discard all occurrences of *Articles* or *Determiners* (1) in the parsed text.

Definite Article - theIndefinite Articles - a, an

Model the remaining parsed text into *unigram*, *bigram*, and *mgram* syntactic symbols according to meaningful *words*, *phrases*, and *clauses*. These symbols are termed *Syntactic Nuggets* (SNs) in this research.

Key language syntactic constructs (SNs) should conform to all language-defined grammar rules derived from the system *Lexicon*, eg.,

noun prefixed with the related *adjective/s*

verb postfixed with the related *adverb/s*

nouns matched with the related *pronouns* occurring in later usage.

subjects > objects > predicates in linguistic precedence

ie., subject dominates the phrase/sentence semantics

object supports the phrase/sentence semantics

predicate indicates nature of association

Semantics Preservation

The SNs in the book title (SN_{BT}s) are matched with the SNs in the book index (SN_Is) and the SNs in book introduction section (SN_{IN}s), ensuring semantic preservation across the three component SNs. This is because the book title presents the key topic of discussion and argumentative thrust, in the chosen text.

SN Notation

SNs can be represented by the following Regular Expression or regex.

(adj*nn | vv adv)** where adjƐ{adjectives}, nnƐ{nouns}, vvƐ{verbs}, advƐ{adverbs}

Process

Compute *Occurrence Probabilities* of each *SN* in the parsed corpus. Key notion is that high probability *SNs* along with their associated (semantically-linked) neighbours provide greater meaning, thrust, and dominate the text's content and discussion.

Occurrence Probability $P(SN_i) = \sum SN_i \; n$, where i = [1.m]

m is number of distinct SNi s in the parsed

n is the size of the parsed corpus

Entropy $E(SN_i) = -\sum P(SN_i) \log_{2\,P}(SN_i)$ ---------- (A) according to (3)

where i denotes the SN index, and

Entropy is measure of Information represented by the particular SN, in bits.

The optimal (x) *Feature* (F) matrix is prepared in accordance with *Features* identified in the *SNs* in the corpus, in ascending order of their respective *Entropies* (E). The empirical value of *x*, the number of rows in the *F* matrix is determined based on the particular experiment and intended synoptic granularity.

The key notion is that low entropy *Features* provide highly informative value in text; in essence dominating the book content in concept and description.

Compose a terse, machine-generated, lexicon-driven, synoptic commentary based upon the tabulated corpus Features and their text-wide significance in description and discussion. Additionally, for volume comparisons, individual text-annotations of a weighted bibliographic factor is effected, based upon the strength of each candidate bibliography.

Significant of the Study

The semantic analysis and synoptic discovery process is fully automated, meaning machine-driven; optimal text comparison/selection being based on text-inherent, rendered *knowledge quality* and *significance*.

Results and Discussion

Consider the simple English sentence "The quick brown fox jumped over the lazy dog". After syntactic recasting, the parsed text would be:

{delArt1] quick brown fox jumped over [delArt2] lazy dog ---(B)

$\boldsymbol{adj_1 adj_2 nn_1}$ $\boldsymbol{vv_1 adv_1}$ $\boldsymbol{adj_3 nn_2}$

Consider the following *syntactically-recasted* corpus

SN_0 = quick brown fox SN_1 = jumped over SN_2 = lazy dog

$SN_{0=}$ quick brown fox SN_3 = leaped over SN_4 = grey wall

where $SN_0 \equiv$ subject, SN_1, $SN_3 \equiv$ predicate, and SN_2, $SN_4 \equiv$ object

Using equation (A) in section 4.4, it can be shown that SN_0 being the principal corpus *subject*, possesses the least Entropy value, thereby presenting the highest potential information in the corpus; the principle property of all referential text themes.

Any descriptive text corpus of *SNs* that *surrounds associatively* with the above recasted phrase (B), will portray primarily the *subject* "quick brown fox", rehearsing its athletic exploits on more sedate objects, such as the "lazy dog".

Automated synoptic annotations can be performed on multiple, parallel texts, enabling synchronous comparisons in terms of the knowledge rendered. Evaluation notifications enable the earnest reader to spend less time searching for referential content, and devote more time to the actual text reading, its comprehension, and the subsequent knowledge application effort, if any.

Conclusion and Recommendations

These machine-generated synopses should be precise, foolproof, and substantiate the inspected piece of literature for usefulness, value, and currency. It should afford seamless analogies with parallel literatures, meaning other text which cover the same topics from other authors, or even different editions of the same text.

The key significance of this research is that implementation is synchronous and fully automated, *prompting* the user with the optimal, machine-endorsed, choice of reference text on demand.

References

Oxford House Language School. (February 2022). How to use Articles in English? Oxford House Language Courses. [Online]. Available: https://oxfordhousebcn.com/en/how-to-use-articles-a-an-the-in-english/

Grammarly Inc. (2023). What are Conjunctions? Definition and Examples. Grammarly. [Online]. Available: https://www.grammarly.com/blog/conjunctions/

Montanar. (November 2007). Introduction to Information Theory. Stanford University Press. [Online]. Available: https://web.stanford.edu/~montanar/RESEARCH/BOOK/partA.pdf