Review On Approaches for Theme Extraction and Sentence Ordering For Prioritization Of Journalistic Notes

Devon Wijesinghe
Department of Computing
Informatics Institute of Technology
Colombo 06, Sri Lanka
devon.2016319@iit.ac.lk

Kaneeka Vidanage
Department of Computing
Informatics Institute of Technology
Colombo 06, Sri Lanka
kaneeka.v@iit.ac.lk

2020 International Conference on Image Processing and Robotics (ICIP) | 978-1-7281-6541-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICIP48927.2020.9367344

***Abstract*— In the stages of pre-writing and writing of a news article, journalists require to process the gathered data to identify important points and events which will predominantly support the main theme of the news story. In relation to the field of computer science, there is a lack of intelligent systems to help organize unstructured journalist data and optimize the news data pre-processing stage. There are existing research projects in the area of natural language processing which are focusing on text ordering and main theme identification of textual documents. However, there is no system, which is fine-tuned for the journalism domain, that can utilize the main theme of an unstructured textual document (journalistic notes) to semantically organize and prioritize text**.

Keywords*— Information-Extraction, Theme-Identification, Semantic-Web, Ontologies, NLP, Sentence-Ordering, Text-Prioritization

I. INTRODUCTION

Journalism is one of the most stressful careers due to tight deadlines, frequent travelling, intensive work conditions and demanding requests of editors [1]. In fact, the profession of a journalist can be found among the top ten stressful jobs around the globe [2]. There is a direct relationship between employee stress levels and performance [3], [4]. High-stress levels of journalists can decrease overall productivity and negatively affect the organization and delay publications. In the process of writing a news article, the "Inverted Pyramid Structure", as illustrated in *figure 1*, is one of the most prominent approaches selected by journalists.

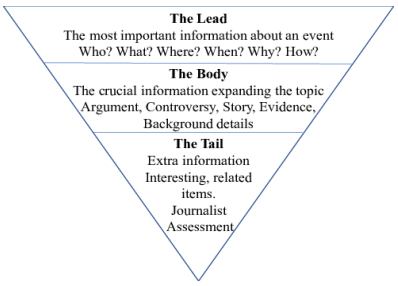

Fig. 1. Inverted pyramid approach [5]

In this approach, since most important and related facts are placed at the beginning of the article, it allows the readers to grab crucial details with less effort [6]. The points place towards the rear are supporting information and will comparatively have less impact on the main theme of the news story. To achieve this, journalists require to manually process the gathered data to identify and prioritize important points and events which will predominantly support the main theme of the news story [7], [8]. To obtain adequate information to construct a particular news story, multiple data sources are used by journalists, which creates a large number of unorganized facts. This will make the pre-processing and prioritizing of important points a daunting task. As a consequence, it will increase the stress levels of journalists and decelerates the speed of producing the final publishable news article.

Furthermore, every news article must convey the "Five W's", which are **Who, What, When, Where, Why** and sometimes **How** [9], [10]. Answers to these questions contribute significantly to the main theme of the article and when we consider the Inverted Pyramid Structure, there is a direct mapping between the placement of news points and the main theme [11]. Hence, it is important to identify the Five W's in the process of crafting a news article.

In relation to the field of computer science, there is a lack of intelligent systems to help identify the main theme from unstructured journalistic data and organize them to optimize the news data pre-processing and writing stages. There are existing research projects in the area of natural language processing which are focusing on text ordering and main theme identification of textual documents. However, there is no system, which is fine-tuned for the journalism domain, that can utilize the main theme of an unstructured textual document (journalistic notes) to semantically organize and prioritize text [12], [13].

II. DOMAIN OVERVIEW

A. Semi-Automated News Writing Approach

Since journalism is a creative job [14] which needs the input from a human. Fully automated news writing, which leverages NLG depend, solely on data and cannot describe new phenomena. Therefore the quality of generated news is not as commendable as compositions created with human involvement. In the semi-automated approach, various software is used as assistive tools in the processes of data