Using Key Phrases to Interpret Semantic Elements of Internet Texts

Duc Huu Pham
http://orcid.org/0000-0003-4564-7986
International University-Vietnam National University Ho Chi Minh City, Vietnam

ABSTRACT

This article discusses the significance of key phrases in unlocking the semantic characteristics of internet texts in obtaining pertinent information from the web. The main issue is with conventional key phrase or keyword-based retrieval systems, which do not account for subtle meaning and contextually conditioned expressions. The solution is to draw upon recent developments in Natural Language Processing (NLP), specifically transformer models such as BERT and GPT, to advance semantic evaluation. Key contributions include the incorporation of NLP techniques to refine document retrieval and semantic interpretation, emphasizing contextual understanding of web texts. Primary findings highlight the potential for NLP models to assist in dealing with challenges such as language uncertainty and cultural divergence and provide more effective and accurate semantic analysis processes for online environments. This article takes forward the development of improved language learners and NLP system tools, enabling them to better interpret and interact with online content that is rich in nature.

INTRODUCTION

The fast expansion of text content online has placed the significance of semantic interpretation at the center of interpreting digital content. This chapter explores how key phrases can help improve semantic interpretation through a more accurate comprehension of complicated texts. The aim of this research is to show how important words, when combined with state-of-the-art semantic analysis methods like topic modeling, contextual word embeddings (like BERT), and knowledge graphs, can help facilitate more insightful understanding of the meaning and context of online texts. Despite the advancement of Natural Language Processing (NLP), semantic analysis is faced with several challenges including the complexity and ambiguity of human language, scalability issues with large data, and context-dependent interpretation. Key phrases play a crucial role in addressing these challenges by serving as indicators of main themes and concepts in texts, thereby enabling more accurate semantic interpretation. This chapter

DOI: 10.4018/407606

will give a thorough evaluation of the proposed methodology using diverse datasets, including tweets, news stories, and product reviews. Through illustrating the strength and scalability of combining key phrases with semantic analysis techniques, this research hopes to push forward the development of more effective language learning tools and natural language processing systems, thus enhancing their abilities to comprehend and interact with complex online materials.

The advancement of technologies has facilitated the use of English as second language or a foreign language to a great extent. The huge resources of knowledge on the Internet have so far assisted language teachers and learners to find many solutions to their problems in language acquisition. To have access to these resources is not an easy task without any appropriate method of exploring documents, specifically, the texts for teaching and learning English. Contemporary systems for document retrieval mainly depend on the use of keywords or key phrases and the commonness of documents in the present time. The most basic form of representing the content of a document is through the use of a list of discrete keywords or key phrases. A document is expressed with words or phrases that are derived from its content. Therefore, this leads to a representation with minimal informational content. Sometimes, the semantic links or the meanings of these keywords or key phrases are not considered. Language users analyzing discourse often face difficulties with key phrase or keyword-based retrieval systems, particularly when trying to express their informational needs through a limited set of descriptive keywords or key phrases and translating these into a query that aligns with the design of the system. This task is especially demanding and challenging for language users with less experience since they may not be able to identify the appropriate keywords for their search queries (Liu, 2021).

Recent NLP methods with transformers have transformed NLP by enabling the use of high-capability tools for deciphering complicated patterns within language such as models like BERT and GPT-4, which have become baselines for activities such as sentiment analysis, text classification, and language generation (Kovács, Csépányi-Fürjes & Tewabe, 2023, October). The aforementioned models employ self-attention mechanisms to pick up contextual cues and word relations, thereby rendering them extremely useful for tasks that involve a profound comprehension of language. These sophisticated approaches to semantic analysis aim at explaining text similarity in transformer models through methods such as layer-wise relevance propagation in an attempt to comprehend feature interaction driving similarity in natural language processing models (Vasileiou & Eberle, 2024). This approach can enhance semantic web page's understanding through more insight into how models interpret and parse languages.

However, in addressing this issue, there is a need to consider applicable ethical issues regarding AI bias. Among the most significant ethical challenges for NLP is bias, which can lead to biased results, based on the data used for model training (Braga, 2023). Thus, if a model is trained with biased data, it has the potential to entrench existing social inequalities. Solutions to counter bias include strict data curation and continuous model checks and bias detection mechanisms. Real-world applications can include sentiment analysis used by businesses to gauge public opinion about products or services, or customer reviews on social media to modify their marketing strategy and improve customer experience. Also, advanced NLP algorithms like GPT-4 have the ability to generate coherent texts, enabling the generation of content for marketing copies, blog posts, reports, and most notably, legal settings to examine legal briefs and contracts to find the main points and forecast upcoming legal rulings based on previous information. This provides time and effort savings and maintains quality. For healthcare, NLP examines clinical records and doctors' notes and helps professionals to make more insightful diagnoses and treatment strategies. For instance, IBM Watson Health uses NLP to process massive amounts of medical data (Patwardhan, Marrone & Sansone, 2023). All the issues above will be theoretically discussed in the following section.