





Grammatical Structure Oriented Automated Approach for Surface Knowledge Extraction from Open Domain Unstructured Text

Muditha Tissera^{1*}  and Ruvan Weerasinghe² 

¹Department of Software Engineering, University of Kelaniya, Kelaniya 11600, Sri Lanka

²School of Computing, University of Colombo, Colombo 00100, Sri Lanka

Abstract

News in the form of web data generates increasingly large amounts of information as unstructured text. The capability of understanding the meaning of news is limited to humans; thus, it causes information overload. This hinders the effective use of embedded knowledge in such texts. Therefore, Automatic Knowledge Extraction (AKE) has now become an integral part of Semantic web and Natural Language Processing (NLP). Although recent literature shows that AKE has progressed, the results are still behind the expectations. This study proposes a method to auto-extract surface knowledge from English news into a machine-interpretable semantic format (triple). The proposed technique was designed using the grammatical structure of the sentence, and 11 original rules were discovered. The initial experiment extracted triples from the Sri Lankan news corpus, of which 83.5% were meaningful. The experiment was extended to the British Broadcasting Corporation (BBC) news dataset to prove its generic nature. This demonstrated a higher meaningful triple extraction rate of 92.6%. These results were validated using the inter-rater agreement method, which guaranteed the high reliability.

Index Terms: Automatic Knowledge Extraction, Relation extraction, Natural Language Processing, Semantic Web, Triples Extraction

I. INTRODUCTION

A. Problem Formation

Open domain also known to as “domain-independent” or “unconstraint-domain” refers to unstructured text from news articles, magazines, World Wide Web (WWW), email text, blogs, and social media comments, where the content is not limited to a single domain. These are the vast information sources among the various types of information generators available today. The knowledge/information facts embedded in these sources are presented using natural language text, which is unstructured and mostly in heterogeneous formats;

thus, only humans can read and understand. However, humans bear limited cognitive processing power, and this never-ending information generation leads to the problem of information overload. Hence, these knowledge sources are not effectively used.

B. Proposed Solution

The main objective of this study is to automatically extract surface knowledge from open-domain news sources and convert it into structured formats so that it can be interpreted by machines. We propose an approach based on the grammatical structure of a sentence to extract triples with a remarkably


Received 20 June 2021, Revised 27 October 2021, Accepted 21 November 2021

*Corresponding Author Muditha Tissera (E-mail: mudithat@kln.ac.lk, Tel: +94-1129-12709)

Department of Software Engineering, University of Kelaniya, Kelaniya 11600, Sri Lanka.

Open Access <https://doi.org/10.6109/jicce.2022.20.2.113>

print ISSN: 2234-8255 online ISSN: 2234-8883

 This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © The Korea Institute of Information and Communication Engineering