HYBRID SYSTEM FOR RECOGNIZING ACRONYM EXPANSIONS USING HEURISTICS AND MACHINE LEARNING TECHNIQUE

A THESIS

Submitted by

MENAHA R

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

**FACULTY OF INFORMATION AND
COMMUNICATION ENGINEERING
ANNA UNIVERSITY
CHENNAI 600 025**

JULY 2021

ABSTRACT

An acronym is a type of abbreviation made up of initial letter or letters of other words. An abbreviation is a short form (SF) of a phrase. The long-form (LF) of an abbreviation is called either a definition or an expansion. Abbreviations and acronyms are commonly used in biomedical literature, scientific and technical articles, information retrieval and web search, etc. Recognizing full forms that are associated with the acronym is important for identifying the meaning of an acronym that facilitates natural language processing and information retrieval from the literature.

Several research works are under practice to automate the recognition of acronym expansion pairs from text and web documents. Heuristics or Machine Learning approaches are prevalently pursued extracting acronym-definition from text or web. Existing heuristics and machine learning approaches recall rate (i.e. Number of retrieved acronym expansion pairs from document rate) is low. Hence, a hybrid model combining heuristics and machine learning is proposed in this work to retrieve more number of acronym expansion pairs from documents. The main objective of the work is to extract abbreviation definition pairs from text documents and also find the list of definitions of the acronym from the web.

Firstly, seven space reduction heuristics are applied to recognize acronyms from the text. Then, three mapping strategies are proposed for doing a sequence labeling task to recognize the expansion of the acronym. Since the usage of acronyms is more in biomedical literature, a biomedical dataset is created from Thalia semantic search engine. Then, the dataset is utilized to identify the potential abbreviation definition pairs. The proposed sequence labeling task gives the features (i.e. labels) of abbreviation definition pairs as output.

Secondly, a single layer neural network (perceptron) is utilized to validate the retrieved definitions using a heuristics approach for improving the accuracy of the proposed system. The features obtained from heuristics-based sequence labeling is given as input to perceptron hence the system is referred to as hybrid model.

Thirdly, the proposed system performance is evaluated using k fold cross-validation method. The standard information retrieval measures such as precision, recall, and F1 are utilized to analyze the system performance in the built-in dataset. Besides, the proposed hybrid model performance is cross-validated in other two publically available datasets such as AB3P, BIOADI, and the results are proved as good.

Finally, the work is extended to find the list of definitions of an acronym from the web for which web-titles are utilized. The proposed heuristics-based sequence labeling is done on web-titles using character-level mapping and word level mapping strategies to extract the definitions. Then, query expansion terms are identified from definitions and the query is reformulated to pull out the list of definitions from the web. The extracted definitions reliability is done through the proposed collocation measure. Lastly, the popularity of the definition is identified through a statistical measure.

The recall rate of proposed hybrid model for finding abbreviation definitions from text is higher than existing system and web based work for finding the list of definition gives more new definitions helps in enriching online dictionary.