Chapter 5
Named Entity Recognition in Document Summarization

Sandhya P.
Vellore Institute of Technology, Chennai Campus, Tamil Nadu, India

Mahek Laxmikant Kantesaria
Vellore Institute of Technology, Chennai Campus, Tamil Nadu, India

ABSTRACT

Named entity recognition (NER) is a subtask of the information extraction. NER system reads the text and highlights the entities. NER will separate different entities according to the project. NER is the process of two steps. The steps are detection of names and classifications of them. The first step is further divided into the segmentation. The second step will consist to choose an ontology which will organize the things categorically. Document summarization is also called automatic summarization. It is a process in which the text document with the help of software will create a summary by selecting the important points of the original text. In this chapter, the authors explain how document summarization is performed using named entity recognition. They discuss about the different types of summarization techniques. They also discuss about how NER works and its applications. The libraries available for NER-based information extraction are explained. They finally explain how NER is applied into document summarization.

DOI: 10.4018/978-1-5225-9373-7.ch005

INTRODUCTION

Named-entity Recognition (NER) is the process in which the entities are extracted for searching, sorting and storing textual information into the categories such as names of organizations, places, persons, expressions of time, quantities or any other measurable quantity. NER system extracts from the plain text in English language or in any other language. NER is also called as entity extraction or entity identification. NER finds the entities from the raw and unstructured data and then define them into different categories. NER reacts differently with different systems. Hence output of one project may not be the same as the output of another project. Although the required outputs of two different systems will be different.

NER is the subtask of the information extraction. It is also a significant component of natural language processing applications. Part-of-Speech tagging, semantic parsers and thematic meaning representations will all outperform when NER is integrated. NER plays a vital role in systems like question answers system, textual entailment, automatic forwarding and news and document searching. NER provides proper and good analytical results. NER is carried out based on different learning methods according to the systems it is being used in. There are three learning methods: Supervised Learning (SL), unsupervised learning (UL) and semi-supervised learning (SSL) (Sekine & Ranchhod, 2007). Supervised learning needs a large dataset. As there is shortage of such datasets, the other two methods are preferred over supervised learning.

Document summarization is a process by which the text is automatically condensed to a summary with the most important information. In general for a human it is required to read the documents and then summarize it. Hence we can extract vital information, we can use them in the use cases such as; dates from feedback system, famous product or model of an item and reviews about the locations. There are many ways to identify the phrases from the text. The simplest method for text identification is by using the dictionary of words.

NER can also be used to process the document. It will extract the words, which are called as entities. These entities will be categorized like persons, organizations, places, time and measurement, and many more. The most important words will then be selected. These words would work as summary for the given document.

In this chapter we explain how document summarization is performed using Named Entity Recognition. First, we discuss about the Named-entity recognition. Then we explain document summarization. The evaluation techniques for text summarization are explained. We then explain how NER works practically with its applications. Then we have mentioned about applying NER to document summarization and issues with it. Then recent advances are explained.