A Highly Accurate PDF-To-Text Conversion System for Academic Papers Using Natural Language Processing Approach

Yong, Tien Fui[1]; Azad, Saiful[2]; Rahman, Mohammed Mostafizur[3]; Zamli, Kamal Z[2]; Rabby, Gollam[2]

[1]Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Bandar Barat, 31900 Kampar, Perak, Malaysia

[2]Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang, 26300 Gambang, Pahang, Malaysia

[3]American International University Bangladesh (AIUB), Banani, Dhaka, Bangladesh

Abstract

Extracting text out of PDF documents is never an easy task when a higher degree of accuracy and consistency are the two main criteria to be attained. Although, there exist a considerable number of such systems; however, most of them are falling short of offering desirable performance especially when academic literature is the concern. Researches, those involved heavily in text mining and project analyzing, need an accurate and consistent supporting tool for PDF-To-Text (PTT) conversion. Therefore, in this paper, we propose a Natural Language Processing based PDF-to-text (NLPDF) conversion system, which comprises of two major steps, namely (i) reads contents from the PDF and (ii) reconstruct the text. The performance of the proposed system is evaluated via four metrics, namely Precision, Recall, *F* -Measure (AF), and standard deviation, and compared with eight other similar benchmarked systems available in the market. The experimental results evidently demonstrate the effectiveness of the proposed system.

Keywords: Edit Distance; Natural Language Processing; PDF-To-Text Conversion