The CSI Journal on
Computing Science and
Information Technology
Vol. 15, No. 1, 2017

Degarbayan: Developing a Persian Paraphrase Corpus by Crowd Sourcing

Reza Maanijou **Seyed Abolghasem Mirroshandel**

Faculty of Engineering, University of Guilan, Rasht, Iran

ABSTRACT

Paraphrase sentences are a different expression of same meanings. Recognizing paraphrase sentences or phrases is an important task in natural language processing systems, but no Persian paraphrase corpus has been developed yet. In this paper, we represent such corpus by using an automatic, unsupervised method for extracting paraphrases. Using data from news agencies and internet news web pages and an algorithm based on Jaccard edit distance, paraphrases are extracted. Paraphrases are extracted in three classes, namely, paraphrase, not paraphrase and irrelevant. Unlike many other approaches, paraphrase phrases are extracted as well as paraphrase sentences. Next, a new crowd sourcing approach based on Telegram messaging robot is used to judge actual labels for each pair of extracted paraphrase candidate. Judged pairs are evaluated and the final corpus is created. Degarbayan corpus consists of 1,523 pairs of paraphrases and the first version of the corpus is available online for academic purposes.

Keywords: Natual Language Processing, Corpus, Crowd Sourcing, Unsupervised Methods, Paraphrase, Distance Measures.