Annals of the Academy of Romanian Scientists
Series on Science and Technology of Information
ISSN 2066 - 2742 Volume **7**, Number **1**/**2014**

AUTOMATIC THREE DIMENSION (3-D) WORD ALIGNMENT APPROACH

Abdel Alnasser ALASFOUR[1], Ştefan TRAUŞAN-MATU[2]

Abstract. *A massive effort is needed to build a parallel aligned corpus, so building a tool to for automatic alignment will be useful for natural language processing in general and information retrieval in particular. In our paper we present a new approach which mixed most of the known alignment techniques to achieve high precision and accuracy ratio without human intervention. A list of most English words was used as anchor list following the Pareto principle.*

Keywords: Alignment, Bi-text, Named Entity, Oracle Text

1. Introduction

Parallel corpora are now one of the most important key resources for multilingual natural language processing including machine learning, information retrieval, and machine translation systems [2]. There are many large scale corpora available offline and online on the WEB. Our concern was to find and build a suitable framework for developing an alignment tool to build any parallel aligned corpus in general and building an Arabic-English parallel corpus in particular. The framework we created is using the available functions and procedures of the "Oracle Text" [1].

Our algorithms were developed in order to be applied directly to any target corpus which will be located in database tables. It gives us the ability to manipulate, analyze and evaluate the results for more accuracy. In order to build such a tool we started by investigating the latest methodologies and approaches in the field of bi-text alignment technologies. In the next sections we will describe in further details each step for achieving our main purpose. We start by teaching our system with the most English used words, keeping in our mind the Pareto principle [14], also known as Pareto law's which says "For many events, roughly 80% of the effects come from 20% of the causes".

Therefore, a list of 1000 common English words was translated to Arabic to be as an initial seed for our bilingual dictionary. This was very useful for developing our alignment tool so that we can align any parallel corpus in the next future.

[1]Ph.D. student, Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Bucharest, Romania, (Nasser_Asfour@yahoo.com).
[2]Member of AOSR. Prof., Ph.D., Faculty of Automatic Control and Computers, University Politehnica of Bucharest; Senior Researcher, Research Institute for Artificial Intelligence of the Romanian Academy, Romania, (stefan.trausan@cs.pub.ro).