نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان – دانشکده مهندسی برق و کامپیوتر

An Introduction to Noor Corpus and its Language Model

Mohammad Hossein Elahimanesh
Islamic Azad University, Qazvin Branch, Qazvin, Iran
Computer Research Center of Islamic Sciences
Qom, Iran
elahimanesh@noornet.net

Behrouz Minaei-Bidgoli
Iran University of Science and Technology, Tehran, Iran
Computer Research Center of Islamic Sciences
Qom, Iran
bminaei@noornet.net

Mohammad Javad Gholami
Computer Research Center of Islamic Sciences
Qom, Iran
mjgholami @noornet.net

Hossein Juzi
Computer Research Center of Islamic Sciences,
Qom, Iran
hjuzi@noornet.net

***Abstract*— In Linguistics, a text corpus is defined as a large group of text documents. Text corpora are used in order to extract the hidden laws of languages. As one application for statistical researches and hidden laws extraction, language models are made to be used for information retrieval applications. In this paper we introduce one of the greatest text corpora in Islamic science which is called Noor Corpus, and then we provide the Language model of this corpus. The Noor Corpus is results of a decade of efforts from theological researchers and computer engineers of Computer Research Center of Islamic Sciences (CRCIS). This corpus includes thousands of Islamic Books are classified into different categories. Most of the existing texts are Arabic and Persian. There are 1.2 billion Arabic words as well as 616 million Persian words. The bigram language models of this corpus have 80 million distinct bigram words in Arabic and 44 million distinct bigram words in Persian.**

Keywords-component; Islamic Corpus; Language Model; Natural Language Processing

I. INTRODUCTION (HEADING 1)

The rapid growth of textual information in the world has led to a huge amount of information whose manage and control seems to be very difficult. To address this problem various techniques are developed by experts in the text mining and information retrieval areas. One of these techniques is applying language models, which is a part of information retrieval science [1].

A large fraction of textual information resources in the world consists of religious resources and we can list several companies and research institutes that develop these resources. Digital libraries, such as the Maktaba Shamila's library[1] or the CRCIS's Noor library[2] , are some of the mentioned resources. This paper introduces one of greatest Islamic dataset that is prepared by the CRCIS. This dataset, known as Noor corpus, contains different fields of Islamic science. The secondary purpose for this paper is to build the language model of this corpus for information retrieval aims. The rest of the paper is organized as follows: in section 2 the corpora that are similar to Noor corpus are criticized. In section 3, we try to define the N-gram language model. Sections 4 and 5, explain Noor corpus statistics along with the results of the language model based on this corpus. Conclusions and future works are presented in section.

II. RELATED WORKS

Many of the previous corpora, in Persian and Arabic, contain newswire text data acquired from Persian and Arabic news sources. The corpus "Arabic Gigaword" whose last edition is called "Arabic Gigaword Fifth Edition" is an example of this type of corpora. This corpus is a huge archive full of newswire texts prepared by Pennsylvania University and Linguist Data Consortium (LDC) and according to the Catalog number LDC2011T11 [4]. In Persian, an instance for these corpora is "Hamshahri". Darrudi et al. have described this corpus with 63 million words (3.97-character average length for each word) [2].

One of the greatest sets of early religious texts can be found in Maktaba Shamila. This program contains more than 2500

This research was supported with CRCIS.
[1] http://shamela.ws
[2] http://www.noorlib.ir