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

Islamic books appearing as a digital library for researchers. Other similar resources can be found in the CRCIS's software programs. Although these resources are very large, but few of them have been used for text mining applications. An example of this application is part of speech tagging of the Holy Quran. Mohamed Elahdj described a statistical part of speech tagger based on Hidden Markov Model for this task [9]. Experimental results of his approach have shown a recognition rate of about 96% on this dataset. Other examples can be found in Al-Hadith classification literatures [3, 5 and 6].

III. STATISTICAL LANGUAGE MODEL

A statistical language model is simply a probability distribution $P(S = w_1 w_2 \dots w_n)$ over all possible sentences[3] [7]. Usage of language models in many of Natural Language Processing (NLP) applications can be founded. Applications such as Part of Speech (POS) tagging, speech recognition and word prediction can be mentioned. There are different types of language models. Two of popular language models are unigram language model and the bigram one. Using of bigram language model is the most frequent model in previous literatures [1].

A. Unigram Language Model

Unigram language model is defined as multiplication of unconditioned probabilities for words of sequence W ($P(W)$). $P(W)$ is computed according to (1):

$$P(W = w_1 w_2 \dots w_n) = \prod_{i=1}^{n} P(w_i) \quad (1)$$

According to this equation the amount of $P(w_i)$ is equal to ratio of number word w_i occurrence to the number whole words in the corpus.

A problem that always appears in language models is that a specific word w_i has never occurred in the corpus. One usual way to smooth this kind of words is to determine their occurrence number as one. We used this smoothing technique for unknown words.

B. Bigram Language Model

In bigram language model, calculation of the probability $P(W)$ is defined as the multiplication of conditional probabilities for w_i s. Conditional probability for each w_i is equal to the probability of w_i given w_{i-1}. Equation (2) shows how to calculate $P(W)$:

$$P(W = w_1 w_2 \dots w_n) = \prod_{i=1}^{n} P(w_i | w_{i-1}) \quad (2)$$

To find $P(w_i|w_{i-1})$ we follow the (3):

$$P(w_i|w_{i-1}) = \frac{f(w_{i-1} w_i)}{f(w_{i-1})} \quad (3)$$

The function f represents the frequency of input string in the corpus.

In this type of model, in addition to smoothing mentioned in last part, we would need another kind of smoothing that is used for unknown bigrams like $w_{k-1}w_k$. Unknown bigrams are those that have not been observed in the corpus. In this paper, we have used linear interpolation technique introduced by Brants [8]. According to this smoothing technique, the way to calculate the probability $P(w_i|w_{i-1})$ is shown in (4):

$$P(w_i|w_{i-1}) = \lambda_1 P(w_i|w_{i-1}) + \lambda_2 P(w_i) \quad (4)$$

Where, $\lambda_1 + \lambda_2 = 1$.

IV. NOOR CORPUS

In this section, we want to introduce Noor corpus which is a corpus produced by the CRCIS. The CRCIS in cooperation with researchers in the fields of computer and religious sciences has made a lot of efforts to digitize Islamic textual documents. The most important aim for this center is to gather and present Islamic texts as desktop programs and websites. These efforts have resulted in many rich libraries in Islamic sciences. In this part, we are going to describe this corpus from different aspects.

A. Statistical reports of corpus

Noor corpus, totally includes 7290 books, each book, by average, has two volumes. The largest book is named "Bihar al-Anwar" consisting of 110 volumes and more than 89 million characters (14 million words). If we count the characters of this corpus regardless of the ones added during the enrichment process, we will have some 8.2 billion characters. Furthermore, the mean length for each book is 1.1 million characters (256,000 words).

B. Corpus's language distribution

The language distribution means the way the textual documents in different languages existing in this corpus, are distributed. Table I shows languages, number of books, characters and words existing in the corpus. In this table the Arabic-Persian books are bilingual and contain texts both in Arabic and Persian.

[3] Or spoken utterances, documents, or any linguistic unit.

C. Corpus's books classification

The researchers of the CRCIS have classified Noor corpus books on the basis of religious categories. Each category in this corpus mostly has caused specific software production in the relative field. Table II explains the distribution of books in this corpus among the most frequent categories.

Distribution of books in Arabic and Persian, in this corpus, is different. Then, table III shows the distribution of Arabic books and table IV, the distribution of Persian books amongst different categories.

TABLE I. NOOR CORPUS LANGUAGE-BASED STATISTICS

Language	*Number of books*	*Number of characters*	*Number of words*
Arabic	4,329	5,250,940,980	1,179,373,254
Persian	2,679	2,624,192,540	616,668,604
Arabic-Persian	231	253,711,374	59,742,677
Others	19	17,056,170	4,045,220
Total	7,290	8,176,706,536	1,867,118,969

TABLE II. NOOR CORPUS STATISTICS ON THE BASIS OF TEXT GENRE

Category	*C(book)*	*C(char)*	*C(word)*
Reasoning-based jurisprudence	695	914,598,202	202,258,740
Legal theory	427	475,000,555	101,542,542
Fatwa-based jurisprudence	206	134,536,929	30,649,561
Geography	158	187,141,263	42,415,480
History	130	179,934,236	41,254,387
Literature	127	109,525,636	26,020,057
Itinerary	117	70,927,900	15,981,599
Nahj al-Balagah	98	144,821,918	33,394,436

TABLE III. ARABIC NOOR CORPUS STATISTICS ON THE BASIS OF TEXT GENRE

Category	*C(book)*	*C(char)*	*C(word)*
Reasoning-based jurisprudence	638	829,267,100	182,198,313
Legal theory	364	397,092,805	83,412,901
Fatwa-based jurisprudence	118	88,093,446	19,599,199
A full collection of historical sources	102	122,949,459	28,063,832
Geography of Cities	89	123,946,653	27,959,309
General translations	73	204,672,655	46,091,378
Peripatecism	66	30,825,401	6,921,714
Exegesis	52	163,688,506	37,044,997

TABLE IV. PERSIAN NOOR CORPUS STATISTICS ON THE BASIS OF TEXT GENRE

Category	*C(book)*	*C(char)*	*C(word)*
Literature	84	52,290,640	12,486,285
Fatwa-based Jurisprudence	84	44,894,915	10,682,714
Hajj	70	21,110,306	4,901,058
Geography of Cities	69	63,194,610	14,456,171
Itinerary	69	45,703,025	10,557,877
Legal theory	61	72,972,086	17,017,638
Qajar Dynasty	59	77,108,236	17,185,921
General History	53	132,957,692	30,747,480

D. Distribution of words length

The mean length for Arabic words in Noor corpus is 3.85 characters and for Persian words, this is 3.55 characters. Figure 1 tells us the way words lengths in Arabic and Persian, in this corpus, is distributed. As you see in figure 1, the most repetitions for Arabic words belong to three-character-length words and for Persian words, two-character-length ones. We notice that any words in the corpus except punctuation Marks are used in this experiment.

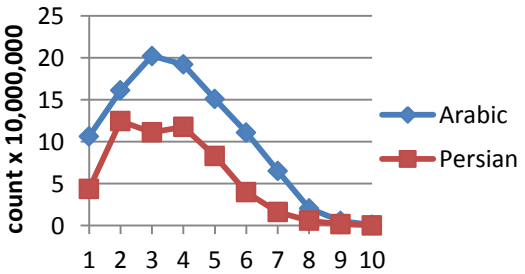

Figure 1. distribution of word length in corpus.

V. NOOR CORPUS'S LANGUAGE MODEL

The result of unigrams and bigrams production is that so far, we have 2,290,525 different unigrams and 79,926,320 different bigrams for Arabic and for Persian, these numbers are, respectively, 1,481,642 and 44,442,394. Production of bigrams is one the most challenging points in preparing the language model for this corpus. Tables number V and VI present ten Arabic and Persian unigrams and bigrams which are repetitious.

TABLE V. TEN ARABIC MOST FREQUENT UNIGRAMS AND BIGRAMS

Id	Ar-unigram		Ar-bigram	
	Word	*Count*	*bigram*	*count*
1	و	92,695,985	، و	22,862,961
2	،	85,659,475	. و	11,278,757
3	.	39,034,326	قال :	3,480,800
4	:	34,336,554	و لا	3,301,256
5	فی	24,970,025	و هو	2,559,540
6	من	23,030,455	: و	2,425,299
7	بن	13,115,677	الله علیه	2,227,760
8	علی	12,170,705	علیه و	2,095,200
9	الله	11,204,098	محمد بن	2,041,946
10	لا	10,957,949	علیه السلام	1,943,612

TABLE VI. TEN PERSIAN REPETITIOUS UNIGRAMS AND BIGRAMS

Id	Pe-unigram		Pe-bigram	
	word	*count*	*bigram*	*count*
1	و	37,594,750	، و	2,641,436
2	،	23,428,992	است .	2,148,170
3	.	21,503,370	. و	1,779,863
4	از	13,292,239	است که	1,648,044
5	که	13,151,501	که در	1,197,099
6	در	12,954,586	و در	1,179,030
7	به	11,755,538	است و	1,140,450
8	را	9,631,841	است ،	1,051,191
9	:	9,494,427	را به	997,047
10	است	8,967,840	و از	866,822

As we discussed in the part for language model, calculating lambda in order to smooth the language model is based on the algorithm offered by Brants [8]. Quantities, coming from the mentioned algorithm, are suitable for both languages, Arabic and Persian, have been put in the Table VII. At the end, Table VIII states ten repetitious characters for each language.

VI. CONCLUSION AND FUTURE WORKS

In this paper we have introduced Noor corpus that in comparison with other Islamic textual corpora, is the greatest one. Number of books, in this corpus, is three times more than the Shamila library. You can compare Noor corpus's number of words with the ones from corpora containing informative sentences.

This corpus and its language model can prepare the information needed for so many of context investigation activities such as part of speech tagging, word prediction and translator machines. For future work, we can discuss more complex language models for this corpus in order to develop more efficient information retrieval.

TABLE VII. LAMBDA QUANTITIES FOR SMOOTHING

	λ_1	λ_2
Arabic	83.8%	16.2%
Persian	82.4%	17.6%

TABLE VIII. TEN MOST FREQUENT CHARACTERS IN NOOR CORPUS

Id	Arabic-char		Persian-char	
	char	*count*	*char*	*count*
1	ا	564,139,203	ا	268,465,150
2	ل	478,263,812	ر	144,880,133
3	م	254,095,481	ن	138,419,184
4	ی	247,568,989	د	130,967,016
5	و	228,604,769	و	123,187,225
6	ن	226,842,132	م	116,182,786
7	ب	166,237,462	ه	111,763,273
8	ه	163,037,900	ی	100,064,001
9	ر	160,472,941	ب	90,631,497
10	ع	144,963,025	ل	78,044,202

REFERENCES

[1] C., Manning, P., Raghavan, H., Schütze, "Introduction to Information Retrieval," Cambridge University Press, 2009.

[2] E., Darrudi, M.R., Hejazi, F., Oroumchian, "Assessment of a modern Persian corpus," Proceedings of the 2nd Workshop on Information Technology & its Disciplines (WITID), ITRC, Iran, 2004.

[3] K., Jbara, "Knowledge Discovery in Al-Hadith Using Text Classification 4Algorithm," Proceeding of 6th Journal of American Science, 2011.

[4] Linguistic Data Consortium, "The Arabic Gigaword corpus," (LDC2011T11), 2011.

[5] M. N., AL-Kabi, S. I., AL-Sinjilawi, "A Comparative Study of the Efficiency of Different Measures to Classify Arabic Text," University of Sharjah Journal of Pure and Applied Sciences, 4(2), pp: 13-26, 2007.

[6] M. N., Al-Kabi, G., Kanaan, R., Al-Shalabi, "Al-Hadith Text Classifier," Proceeding of 5th Journal of Applied Sciences, pp: 584-587, 2005.

[7] R., Rosenfeld, "Two Decades of Statistical Language Modeling: Where Do We Go From Here?," Proceeding of the IEEE, 88(8), 2000.

[8] T., Brants, "TnT: A statistical part of speech tagger," Proceedings of the 6th Conference on Applied Natural Language Processing, Apr. 29-May 04, Association for Computational Linguistics Morristown, New Jersey, USA, 2000.

[9] Y. O. M., Elhadj, "Statistical Part-Of-Speech Tagger for Traditional Arabic Texts," Proceeding of 5th Journal of Computer Sciences, V 11, pp: 794-800, 2009.