Exploring Arabic text diacritization approaches in view of establishing an action plan for developing an open source diacritizer*

Dr. Nada Ghneim ** **Dr. Ghaida Rebdawi** ***

Abstract

The absence of diacritization in Arabic texts is one of the most important challenges facing the automatic Arabic Language processing. When reading, Arabic reader can expect the correct diacritics of words, while computers need algorithms to restore the diacritization based on knowledge of different levels. Diacritization here includes all the diacritics (dama, fatha, kasra, sokon), in addition to alshadda, and altanween.

Some diacritization methods are based on the linguistic processing of texts, while other methods are based on statistical methods using textual corpus. Some systems integrate the two methodologies in hybrid approaches.

In this paper we present a comprehensive study of different methods that have been adopted in these diacritization systems. In addition, we review the various corpuses that have been used for tests and evaluation, then suggest the specifications of the Arabic corpus needed for diacritization systems, and the standards that the evaluation process must take into consideration. The main objective is to develop an action plan for the construction of an automatic diacritizer of Arabic texts under the auspices of ALECSO, with the participation of many research entities from different countries.

Keywords: automatic Arabic language processing, automatic diacritization of Arabic texts, morphological analysis, evaluation corpus, diacritizaters evaluation.

* For The paper in Arabic see pages (277-294)

*** Informatics Department, HIAST, Damascus, Syria

** Informatics Department, HIAST, Damascus, Syria

References:

[1] N. Habash, O. Rambow, 2007, "Arabic Diacritization through Full Morphological Tagging", Proceedings of 8th Meeting of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies Conference.

[2] M. Rashwan, M. Al-Badrashiny, M. Attia and S. M. Abdou, 2009, "A Hybrid System for Automatic Arabic Diacritization", Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, Cairo, Egypt, April 2009.

[3] M. Maamouri, A. Bies, and T. Buckwalter. 2004. The Penn Arabic Treebank: Building a large-scale annotated arabic corpus. In NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt.

[4] R. Nelken and S. M. Shieber. 2005. Arabic Diacritization Using Weighted Finite-State Transducers. In Proc. of the ACL 2005 Workshop On Computational Approaches To Semitic Languages, Ann Arbor, Michigan, USA.

[5] M. Elshafei, H. Almuhtasib and M. Alghamdi, 2006, "Machine Generation of Arabic Diacritical Marks", The 2006 World Congress in Computer Science Computer Engineering, & Applied Computing . Las Vegas, USA.

[6] J. Giménez and L. Márquez, 2004, SVMTool: A general POS tagger generator based on Support Vector Machines. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal.

[7] I. Zitouni, J. S. Sorensen, and R. Sarikaya. 2006. Maximum Entropy Based Restoration of Arabic Diacritics. In Proc. of the 4th Annual Meeting of ACL, Australia.

[8] M. Attia, M. Rashwan, 2004, A Large-Scale Arabic PoS Tagger Based on a Compact Arabic PoS Tags Set, and Application on the Statistical Inference of Syntactic Diacritics of Arabic Text Words, Proc. of the Arabic Language Technologies & Resources Int'l Conf.; NEMLAR, Cairo.

[9] H. Safadi., O. Al Dakkak and N. Ghneim, 2006, "Computational Methods to Vocalize Arabic Texts" 2nd Workshop W3C, Herakilon, Greece.

[10] A. Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In Conf. on Empirical Methods in NLP.

[11] O. Emam and V. Fischer. 2005. Hierarchical Approach for the Statistical Vowelization of Arabic Text. Technical report, IBM Corporation Intellectual Property Law, Austin, TX, US.

[12] T. Schlippe, T. Nguyen, and S. Vogel, 2008, "Diacritization as a Machine Translation Problem and as a sequence Labeling Problem", The 8th Conference of the Association for Machine Translation in the Americas, Waikiki, Hawaii, 270-278.

[13] M.S. Ryan, and G.R. Nudd, 1993, The Viterbi algorithm, Technical Report. Department of Computer Science, Coventry, UK.

[14] A. Mazroui, A. Mezian, A. Lkhwaja, M. weld Bebah, A. Bodlal, 2010, "A Statistical Approach for Arabic Diacritization", Enriching Arabic Digital Content Workshop, in Arabic, Damascus-Syria.