AN AUTOMATIC DIACRITIZATION ALGORITHM FOR UNDIACRITIZED ARABIC TEXT

AYMAN AHMAD MOHAMMAD ZAYYAN

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)

UNIVERSITI UTARA MALAYSIA

2017

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

AYMAN AHMAD MOHAMMAD ZAYYAN

calon untuk Ijazah
(candidate for the degree of) MASTER OF SCIENCE (INFORMATION TECHNOLOGY)

telah mengemukakan tesis / disertasi yang bertajuk:
(has presented his/her thesis / dissertation of the following title):

"AN AUTOMATIC DIACRITIZATION ALGORITHM FOR UNDIACRITIZED ARABIC TEXT"

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : ***17 Mei, 2017.***
That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on: ***May 17, 2017.***

Pengerusi Viva: *(Chairman for VIVA)*	Assoc. Prof. Dr. Osman Ghazali	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	Assoc. Prof. Dr. Akram M Z M Khedher	Tandatangan *(Signature)*
Pemeriksa Dalam: *(Internal Examiner)*	Dr. Samry @ Mohd Shamrie Sainin	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	Dr. Husniza Husni	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	Dr. Shahrul Azmi Mohd Yusof	Tandatangan *(Signature)*

Tarikh:
(Date) May 17, 2017

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences

UUM College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

Abstrak

Bahasa Arab Standard Moden (MSA) digunakan hari ini dalam kebanyakan media bertulis dan beberapa media pertuturan. Ia bagaimanapun, bukan dialek asal mana-mana negara. Kebanyakan teks ini telah ditulis dalam dialek Mesir, kerana ia dianggap dialek yang paling banyak digunakan dan difahami di seluruh Timur Tengah. Seperti Bahasa Semitik lain, dalam Bahasa Arab bertulis, vokal pendek tidak ditulis tetapi diwakili dengan tanda diakritik. Walau bagaimanapun, tanda ini tidak digunakan dalam kebanyakan teks bahasa Arab moden (buku, akhbar, dll.). Ketiadaan tanda diakritik mewujudkan kekaburan yang besar kerana perkataan yang tidak bertanda diakritik mungkin bersesuaian dengan lebih daripada satu bentuk *diacritization* yang betul (*vowelization*). Oleh itu, matlamat penyelidikan ini adalah untuk mengurangkan kekaburan ketiadaan tanda diakritik menggunakan algoritma hibrid dengan ketepatan yang lebih tinggi berbanding sistem terkini bagi MSA. Selain itu, kajian ini juga adalah untuk melaksanakan dan menilai ketepatan algoritma untuk teks Bahasa Arab dialek. Reka bentuk algoritma yang dicadangkan berdasarkan dua teknik utama seperti berikut: statistik n-gram bersama dengan anggaran kebarangkalian maksimum dan penganalisis morfologi. Menggabungkan perkataan, morfem, dan aras huruf serta sub-model mereka bersama-sama ke dalam satu platform untuk meningkatkan ketepatan *diacritization* automatik adalah cadangan penyelidikan ini. Selain itu, dengan menggunakan ciri *case ending diacritization*, iaitu mengabaikan tanda diakritik pada huruf terakhir perkataan, menunjukkan peningkatan signifikan terhadap penambahbaikan ke atas ralat. Sebab peningkatan yang luar biasa ini adalah bahawa Bahasa Arab melarang menambah tanda diakritik terhadap beberapa huruf. Algoritma yang dicadangkan menunjukkan prestasi yang baik sebanyak 97.9% apabila digunakan untuk korpora MSA (Tashkeela), 97.1% apabila diaplikasikan pada LDC's Arabic Treebank-Part 3 v1.0 dan 91.8% apabila digunakan bagi korpus dialektal Mesir (CallHome). Sumbangan utama penyelidikan ini ialah algoritma hibrid untuk *diacritization* automatik teks MSA yang tiada diakritik dan teks Bahasa Arab dialek. Algoritma yang dicadangkan digunakan dan dinilai pada dialek Bahasa harian Mesir, dialek yang paling luas difahami dan digunakan di seluruh dunia Arab yang dianggap sebagai kali pertama berdasarkan kajian literature.

Kata kunci: Diacritization automatik, tanda diakritik, penganalisis morfologi, Anggaran kebarangkalian maksimum, statistic n-gram.

Abstract

Modern Standard Arabic (MSA) is used today in most written and some spoken media. It is, however, not the native dialect of any country. Recently, the rate of the written dialectal Arabic text increased dramatically. Most of these texts have been written in the Egyptian dialectal, as it is considered the most widely used dialect and understandable throughout the Middle East. Like other Semitic languages, in written Arabic, short vowels are not written, but are represented by diacritic marks. Nonetheless, these marks are not used in most of the modern Arabic texts (for example books and newspapers). The absence of diacritic marks creates a huge ambiguity, as the un-diacritized word may correspond to more than one correct diacritization (vowelization) form. Hence, the aim of this research is to reduce the ambiguity of the absences of diacritic marks using hybrid algorithm with significantly higher accuracy than the state-of-the-art systems for MSA. Moreover, this research is to implement and evaluate the accuracy of the algorithm for dialectal Arabic text. The design of the proposed algorithm based on two main techniques as follows: statistical n-gram along with maximum likelihood estimation and morphological analyzer. Merging the word, morpheme, and letter levels with their sub-models together into one platform in order to improve the automatic diacritization accuracy is the proposition of this research. Moreover, by utilizing the feature of the case ending diacritization, which is ignoring the diacritic mark on the last letter of the word, shows a significant error improvement. The reason for this remarkable improvement is that the Arabic language prohibits adding diacritic marks over some letters. The hybrid algorithm demonstrated a good performance of 97.9% when applied to MSA corpora (Tashkeela), 97.1% when applied on LDC's Arabic Treebank-Part 3 v1.0 and 91.8% when applied to Egyptian dialectal corpus (CallHome). The main contribution of this research is the hybrid algorithm for automatic diacritization of undiacritized MSA text and dialectal Arabic text. The proposed algorithm applied and evaluated on Egyptian colloquial dialect, the most widely dialect understood and used throughout the Arab world, which is considered as first time based on the literature review.

Keywords: Automatic diacritization, Diacritic marks, morphological analyzer, maximum likelihood estimation, statistical n-gram.

Acknowledgement

All praise is due to Allah, who guided me to this.

I would like to express my sincere gratitude to my supervisor; Dr. Husniza binti Husni and Dr. Shahrul Azmi Mohd Yusof. I'm greatly indebted to their assistance, guidance and support.

I would like to thank the Arabic language expert Dr. Mohamed Elmahdy, German University in Cairo, for his generous help.

I am very grateful for my dear parents, wife, daughter and my friends whom I consider as my brothers. Thank you all for being always there when I needed you most. Thank you for believing in me and supporting me. I believe that without your support and your prayers, none of this work would be accomplished. Finally, I hope this thesis be a useful addition to the research activities of Arabic natural language processing.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

1- **MSA:** Modern Standard Arabic.

2- **OOV:** Out of Vocabulary.

3- **WER1:** Word Error Rate, without considering the case ending.

4- **WER2:** Word Error Rate, with considering the case ending.

5- **DER1:** Diacritization Error Rate, without considering the case ending.

6- **DER2:** Diacritization Error Rate, with considering the case ending.

CHAPTER ONE
INTRODUCTION

1.1 Background

Arabic is the largest still living Semitic language in terms of number of speakers that exceeds 350 million [1]. Arabic is natively spoken by people in the Middle East as well as for religious texts by Muslims in many countries. Modern Standard Arabic (MSA) [2] is the form of Arabic closest to the classical Arabic used in the Qur'an and other ancient texts. MSA is used today in most written and some spoken media. It is, however, not the native dialect of any country. Recently the rate of the written dialectal Arabic text increased dramatically. It is being used as a daily life language communication and for expressing the ideas across the World Wide Web [3]. Most of these texts have been written in the Egyptian dialectal, as it is considered the most widely dialect used and understood throughout the Middle East [3]. Moreover, due to the limited availability of the dialectal data. Like other Semitic languages, in written Arabic, short vowels are not written, but are represented by diacritic marks. Nonetheless, these marks are not used in most of the modern Arabic texts (books, newspapers, etc).

The Arabic language is one of the languages where the intended pronunciation of a certain word cannot be fully determined by its standard orthographic representation. Therefore, a set of special diacritic marks is needed in order to indicate the intended correct pronunciation, see Table 1.1.

The contents of the thesis is for internal user only

REFERENCES

[1] M. Rashwan, A. Al Sallab, H. Raafat and A. Rafea, "Deep Learning Framework with Confused Sub-Set Resolution Architecture for Automatic Arabic Diacritization," *IEEE/ACM Transactions On Audio, Speech, And Language Processing,* vol. 23, no. 3, pp. 505-516, 2015.

[2] G. Abandah, A. Graves and B. Al-Shag, "Automatic diacritization of Arabic text using recurrent neural networks," *International Journal on Document Analysis and Recognition (IJDAR),* vol. 18, no. 2, pp. 183-197, 2015.

[3] H. Abo Bakr, K. Shaalan and I. Ziedan, "A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic," in *The 6th international conference on informatics and systems, infos2008*, Cairo, Egypt, 2008.

[4] S. Harrat , M. Abbas , K. Meftouh and K. Smaïli, "Diacritics Restoration for Arabic Dialects," in *14th Annual Conference of the International Speech Communication Association* , Lyon, France, 2013.

[5] A. Said, M. El-Sharqwi, A. Chalabi and E. Kamal, "A Hybrid Approach for Arabic Diacritization," in *18th International Conference on Applications of Natural Language to Information Systems*, Salford, UK, 2013.

[6] A. Azmi and R. Almajed, "A survey of automatic Arabic diacritization techniques," *Natural Language Engineering,* vol. 21, no. 3, pp. 477-495, 2013.

[7] A. Shahrour, S. Khalifa and N. Habash, "Improving Arabic Diacritization through Syntactic Analysis," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015.

[8] M. Rashwan, A. Al Sallab, H. Raafat and A. Rafea, "Automatic Arabic diacritics restoration based on deep nets," in *Empirical Methods in Natural Language Processing*, Doha - Qatar, 2014.

[9] Y. Hifny, "Restoration of Arabic Diacritics using Dynamic Programming," in *8th International Conference on Computer Engineering & Systems (ICCES)*, Cairo, Egypt, 2013.

[10] A. Al-Taani and S. Abu Al-Rub, "A Rule-Based Approach for Tagging Non-Vocalized Arabic Words," *The International Arab Journal of Information Technology,* vol. 6, no. 3, pp. 320-328, 2009.

[11] M. Ameur, Y. Moulahoum and A. Guessoum, "Restoration of Arabic Diacritics Using a Multilevel Statistical Model," *Springer International Publishing,* vol. 456, pp. 181-192, 2015.

[12] R. Nelken and S. Shieber, "Arabic diacritization using weighted finite-state transducers," in *In Proceedings of the 2005 ACL Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, Michigan, 2005.

[13] M. Bebah, C. Amine, M. Azzeddine and L. Abdelhak, "Hybrid Approaches For Automatic Vowelization Of Arabic Texts," *International Journal on Natural Language Computing,* vol. 3, no. 4, pp. 53-71, 2014.

[14] N. Habash and O. Rambow, "Arabic Diacritization through Full Morphological Tagging," in *Proceedings of The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, 2007.

[15] Y. Hifny, "Higher Order n-gram Language Models for Arabic Diacritics Restoration," in *The Twelfth Conference on Language Engineering*, Cairo, Egypt, 2012.

[16] Y. Hifny, "Smoothing Techniques for Arabic Diacritics Restoration," in *The Twelfth Conference on Language Engineering*, Cairo, Egypt, 2012.

[17] M. Alghamdi, Z. Muzaffar and H. Alhakami, "Automatic Restoration Of Arabic Diacritics: A Simple, Purely Statistical Approach," *The Arabian Journal for Science and Engineering,* vol. 35, no. 2C, pp. 125-135, 2010.

[18] M. Alghamdi and Z. Muzafar, "KACST Arabic Diacritizer," in *the First International Symposium on Computers and Arabic Language*, Riyadh, Saudi Arabia, 2007.

[19] M. Elshafei, H. Al-Muhtaseb and M. Alghamdi, "Statistical Methods for Automatic diacritization of Arabic text," in *Saudi 18th National Computer Conference*, Riyadh, Saudi Arabia, 2006.

[20] M. Elshafei, H. Al-Muhtaseb and M. Alghamdi, "Machine Generation Of Arabic Diacritical Marks," in *The 2006 International Conference on Machine Learning; Models, Technologies & Applications*, Las Vegas, Nevada, USA, 2006.

[21] I. Zitouni, J. Sorensen and R. Sarikaya, "Maximum Entropy Based Restoration of Arabic Diacritics," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, Sydney, Australia, 2006.

[22] S. Ananthakrishnan, S. Bangalore and S. Narayanan, "Automatic Diacritization of Arabic Transcripts for Automatic Speech Recognition," in *In Proceedings of the International*

Conference on Natural Language Processing (ICON-05), Kanpur, India, 2005.

[23] Y. Gal, "An hmm approach to vowel restoration in arabic and hebrew," in *Workshop on Computational Approaches to Semitic Languages*, Philadelphia, USA, 2002.

[24] K. Shaalan, H. M Abo Bakr and I. Ziedan, "A Hybrid Approach for Building Arabic Diacritizer," in *The Proceedings of the 12th European Chapter of the Association for Computational Linguistics (EACL 2009) Workshop on Computational Approaches to Semitic Languages*, Athens, Greece, 2009.

[25] A. Said, M. El-Sharqwi, A. Chalabi and E. Kamal, "A Hybrid Approach for Arabic Diacritization," in *18th International Conference on Applications of Natural Language to Information Systems*, Salford, UK, 2013.

[26] M. Rashwan, M. Al-Badrashiny, M. Attia, S. Abdou and A. Rafea, "A Stochastic Arabic Diacritizer Based on a Hybrid of Factorized and Unfactorized Textual Features," *IEEE Transactions On Audio, Speech, And Language Processing,* vol. 19, no. 1, pp. 166-175, 2011.

[27] T. Schlippe, T. Nguyen and S. Vogel, "Diacritization as a Machine Translation Problem and as a Sequence Labeling Problem," in *The Eighth Conference of the Association for Machine Translation in the Americas - AMTA 2008*, Hawaii, 2008.

[28] M. Rashwan, M. Elbadrashiny, M. Attia and S. Mahdy Abdou, "A hybrid system for automatic arabic diacritization," in *The 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, 2009.

[29] A. El-Desoky, R. Schluter and H. Ney, "A Hybrid Morphologically Decomposed Factored Language Models for Arabic LVCSR," in *The 2010 Annual Conference of the North American Chapter of the ACL*, Los Angeles, California, 2010.

[30] M. Rashwan, M. Al Badrashiny, M. Attia, S. Abdou and A. Rafea, "Stochastic Arabic hybrid diacritizer," in *Natural Language Processing and Knowledge Engineering, 2009*, Dalian, China, 2009.

[31] N. Habash and O. Rambow, "Arabic Diacritization through Full Morphological Tagging," in *The Annual Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, Rochester, New York, 2007.

[32] T. Zerrouki, "Arabic corpora resources," Tashkeela collection from the Arabic Al-Shamela library, 19 July 2011. [Online]. Available: http://aracorpus.e3rab.com. [Accessed 27 November 2014].

[33] "Linguistic Data Consortium," LDC, [Online]. Available: https://www.ldc.upenn.edu. [Accessed 17 March 2016].

[34] "Linguistic Data Consortium," LDC, [Online]. Available: https://www.ldc.upenn.edu/. [Accessed 7 May 2016].

[35] S. Ross, Introductory Statistics, 3rd edition, Academic Press, 2005.

[36] D. Jurafsky and J. H. Martin, Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, New Jersey: Pearson Education, 2009.

[37] "Catalogue of Language Resources," European Land Registry Association (ELRA), [Online]. Available: http://catalog.elra.info/index.php?language=en. [Accessed 17 12 2015].

[38] A. Chennoufi, A. Mazroui and A. Lakhouaja, "HYBRID APPROACHES FOR AUTOMATIC," *International Journal on Natural Language Computing,* vol. 3, no. 4, pp. 53-71, 2014.

[39] E. Kamal, A. Said, M. El-Sharqwi and A. Chalabi, "A Hybrid Approach for Arabic Diacritization," in *18th International Conference on Applications of Natural Language to Information Systems*, Salford, UK, 2013.

[40] M. A. Rashwan, M. Elbadrashiny and S. Mahdy Abdou, "A Hybrid System for Automatic Arabic Diacritization," in *The 2nd International Conference on Arabic Language Resources and Tools.*, Cairo, Egypt., 2009.

[41] K. Shaalan, H. M Abo Bakr and I. Ziedan, "A Hybrid Approach for Building Arabic Diacritizer," in *The Proceedings of the 12th European Chapter of the Association for Computational Linguistics*, Athens, Greece, 2009.

[42] M. Bebah, C. Amine, M. Azzeddine and L. Abdelhak, "Hybrid Approaches For Automatic Vowelization Of Arabic Texts," *International Journal on Natural Language Computing (IJNLC),* vol. 3, no. 4, pp. 53-71, 2014.