Arabic Information Retrieval

Kareem Darwish
Qatar Computing Research Institute
kdarwish@qf.org.qa

Walid Magdy
Qatar Computing Research Institute
wmagdy@qf.org.qa

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

K. Darwish and W. Magdy. *Arabic Information Retrieval*. Foundations and Trends® in Information Retrieval, vol. 7, no. 4, pp. 239–342, 2013.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-60198-777-8

Foundations and Trends® in Information Retrieval
Volume 7, Issue 4, 2013

Editorial Board

Full text available at: http://dx.doi.org/10.1561/1500000031

Editorial Scope

Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation, and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2013, Volume 7, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in Information Retrieval
Vol. 7, No. 4 (2013) 239–342

DOI: 10.1561/1500000031

now
the essence of knowledge

Arabic Information Retrieval

Kareem Darwish
Qatar Computing Research Institute
kdarwish@qf.org.qa

Walid Magdy
Qatar Computing Research Institute
wmagdy@qf.org.qa

Contents

Abstract

In the past several years, Arabic Information Retrieval (IR) has garnered significant attention. The main research interests have focused on retrieval of formal language, mostly in the news domain, with ad hoc retrieval, OCR document retrieval, and cross-language retrieval. The literature on other aspects of retrieval continues to be sparse or non-existent, though some of these aspects have been investigated by industry. Others aspects of Arabic retrieval that have received attention include document image retrieval, speech search, social media and web search, and filtering. However, efforts on different aspects of Arabic retrieval continue to be deficient and severely lacking behind efforts in other languages. The survey covers: 1) general properties of the Arabic language; 2) some of the aspects of Arabic that affect retrieval; 3) Arabic processing necessary for effective Arabic retrieval; 4) Arabic retrieval in public IR evaluations; 5) specialized retrieval problems, namely Arabic-English CLIR, Arabic Document Image Retrieval, Arabic Social Search, Arabic Web Search, Question Answering, Image retrieval, and Arabic Speech Search; 6) Arabic IR and NLP resources; and 7) open IR problems that require further attention.

K. Darwish and W. Magdy. *Arabic Information Retrieval*. Foundations and Trends® in Information Retrieval, vol. 7, no. 4, pp. 239–342, 2013.

DOI: 10.1561/1500000031.

1

Introduction

Most early studies on Arabic IR relied on relatively small test collections containing hundreds of documents that are composed of character-coded Arabic text (7; 13; 88). Increased interest in Arabic processing and retrieval in the early 2000's led to significant work that mostly relied on a single large collection (from TREC-2001/2002) (68; 129). However, most of the work was restricted to ad hoc retrieval and cross-language retrieval. Later work focused on other aspects of Arabic retrieval including document image retrieval, speech search, social media and web search, and filtering. However, efforts on different aspects of Arabic retrieval continue to be deficient and severely lacking behind efforts in other languages. This survey reviews recent literature pertaining to different aspects of Arabic IR including different domains and applications. It also describes some of the Arabic specific challenges affecting retrieval and some of the proposed solutions to these challenges. Further, it identifies the available resources and open areas of research to aid those interested in Arabic IR research.

The remainder of this introductory section presents general interesting aspects of Arabic and outlines the content of subsequent sections in the survey.

1.1 The Arabic Language

Arabic is the most widely spoken Semitic language with an estimated 400 million speakers. Arabic shares many commonalities with other Semitic languages. These commonalities pertain to morphology, vocabulary, word order (subject-verb-object and verb-subject-object), use of short and long vowels, etc. For example, Arabic and Hebrew words are typically derived from roots that are composed of two, three, or four letters, with three letter (triliteral) roots being the most common. Words are constructed from roots by possibly inserting infixes, adding prefixes and suffixes, or doubling constants. Diacritics, which are often omitted in writing, help disambiguate words. Nouns can be singular, plural, or dual, and masculine or feminine.

Arabic has a broad sphere of influence which is mostly due to: a) religious reasons, where Arabic is the language of Islamic scholarship and that of the Muslim holy book, the Qur'an; and b) Arabic was the language of science and technology during the Middle Ages, with major Arabic universities in Spain, Africa, and the Middle East being learning hubs. Consequently, Arabic is part of school curricula in most majority non-Arab Muslim countries such as Iran and Pakistan. Arabic is also an official language in other countries such as Eritrea, Chad, and Somalia. Arabic had influence, mostly in terms of vocabulary, on many other languages such as Spanish, Turkish, Persian, Urdu, Swahili, and Hausa. Further, Arabic script is used for writing many languages such as Persian, Urdu, Kurdish, Pashto, and Dari.

The Arab population is generally a young population with an average age in the Arab World slightly less than 24.[1] The Arabic language is ranked as the seventh top language on the web.[2] The Arab Internet users constitute 3.3% of the Internet users worldwide. Although Arabic is ranked seventh among languages on the web, it is the fastest growing language on the web among all other languages (Figure 1.1). The number of Arab Internet users grew from 2.5 million in 2000 to 65 million in 2011. Internet penetration in the Arab World is estimated to

[1] `https://www.cia.gov/library/publications/the-world-factbook/index.html`

[2] `http://www.internetworldstats.com/`

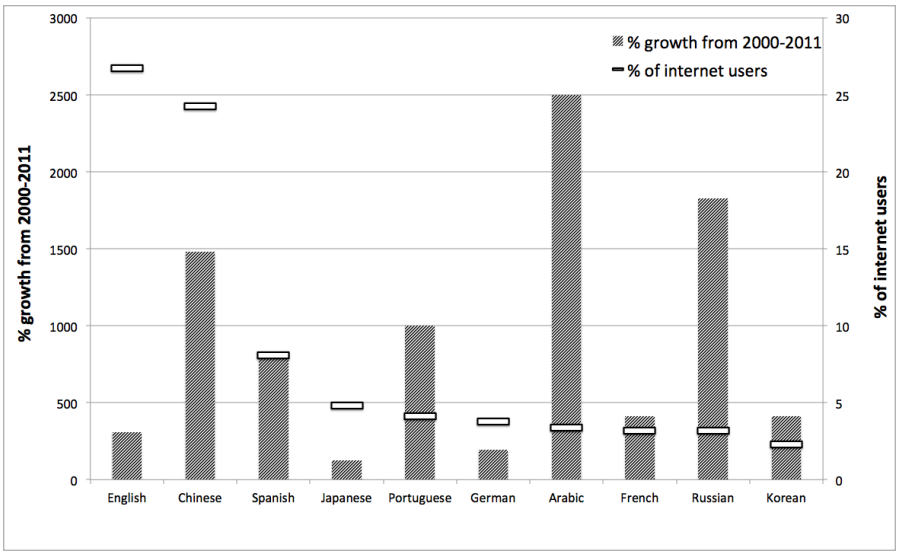

Figure 1.1: Top 10 languages in the Internet by 31 Dec 2011 (www.internetworldstats.com)

be 24%, which is lower than the global average of 32.7%. There are an estimated 45 million Arab Facebook users constituting roughly 5.6% of Facebook users globally. Though no exact estimates are available, Arabic online content is believed to constitute less than 1.5% of the global content. The relative size of Arabic forum content is disproportionately larger compared to the English forum content. English forum content is often considered of lower quality. However, such content is often of high quality in Arabic.

Modern Standard Arabic (MSA) is the lingua franca for the so-called Arab world, which includes northern Africa, the Arabian Peninsula, and Mesopotamia. Figure 1.2 shows a sample document written in MSA, which is an article from the Aljazeera.net news website. The article is written in MSA and would generally be understood by most Arabic speakers. However, Arabic speakers generally use dramatically different languages (or dialects) in daily interactions. There are six dominant dialects, which are Egyptian (85+ million speakers), Maghrebi (75+ million), Levantine (35+ million), Iraqi (25 million), Gulf (25+

تناولت الصحافة الأميركية والبريطانية تداعيات الأحداث في مصر واتفاق السلام الإسرائيلي الفلسطيني، فأشارت مجلة تايم إلى تقرير مصور عن مجزرة رابعة العدوية وتحدثت صحيفة غارديان عن عودة مصر لوحدات الشرطة السرية، بينما ركزت صحيفة ديلي تلغراف على أسباب استعداد إسرائيل لدراسة اتفاقية السلام.

وفي الشأن المصري تناولت مجلة تايم الأميركية تقريرا مصورا لشاب يدعى مصعب الشامي يعمل كصحفي مستقل وهو يدرس في كلية الصيدلة. وعندما بدأت الشرطة تطلق النار على المتظاهرين فيميدان رابعة العدوية بالقاهرة وثق الشامي آثار المجزرة وأرسلها إلى عدة وسائل إعلام عالمية منها مجلة تايم.

ويقول مصعب الشامي إنه علم بالأمر من شبكة التواصل الاجتماعي الساعة ١.٣٠ صباح يوم ٢٧ يوليو/تموز وكان وقتها في وسط القاهرة ولم تكن مفاجأة له لأنه كان يتوقع نوعا من العنف في أي وقت قريب. ويضيف أنه وصل إلى مسرح الأحداث بعد الثانية صباحا وكان الجو ملبدا بالغازات المدمعة، وكان يقف خلف خطوط الشرطة حيث كان أفراد الأمن المركزي يواجهون مؤيدي مرسي، وقد منعته الشرطة هو وصحفيين آخرين من الاقتراب لكنه تمكن من رؤية ضباط الشرطة ومعهم أناس في ملابس مدنية يشتبكون مع شباب الإخوان المسلمين، وكان في أيدي مالا يقل عن اثنين من المدنيين مسدسات.

Figure 1.2: Sample Arabic document from Aljazeera news website

million), and Yemeni (20+ million).[3] Aside from those, there are tens of different Arabic dialects along with variations within the dialects. Due to the spread of social media, users are increasingly using Arabic dialects online. These dialects may differ in vocabulary, morphology, and spelling from MSA and most do not have standard spellings.

The fast growth of Arabic content on the web and the large variations between MSA and different dialects make it essential to develop effective IR systems. Figure 1.1 shows the number of users for some of the languages and their growth trends. In this survey, the efforts exerted for developing these systems are explored for different IR applications.

[3] http://en.wikipedia.org/wiki/Varieties_of_Arabic

1.2 The Remainder of the Survey

The subsequent sections cover the following topics:

Section 2, entitled "Arabic Features Affecting Retrieval", presents key aspects of Arabic that affect retrieval. These key aspects include Arabic orthography and morphology; the use of MSA vs. dialects; the differences between formal and informal text; the use of non-standard textual representations; and Arabic properties affecting the retrieval of content in different modalities, namely print and speech.

Section 3, entitled "Arabic Preprocessing and Indexing", presents the core preprocessing steps that are required to prepare Arabic text for effective IR. The preprocessing steps including handling different encodings of Arabic, orthography, morphology, lexical and spelling variations, and stopwords. It also introduces effective index terms for Arabic.

Section 4, entitled "Arabic IR in Shared-Task Evaluations", explores the presence of the Arabic language in different IR evaluation campaigns such as TREC, TDT, BOLT, and CLEF. It also presents the different IR tasks at the campaigns, namely ad hoc retrieval, filtering, cross-language retrieval, topic detection and tracking, and question answering.

Section 5, entitled "Domain-specific IR", surveys work on different IR applications. These applications include cross-language IR, document image retrieval, general web search, social search, question answering, image retrieval, and speech search. The section addresses some of the challenges associated with different applications and some of the solutions that are reported in the literature.

Section 6, entitled "Open Research Areas in Arabic IR", explores open areas of research that require more work. These areas include ad hoc IR, question answering, social search, and web search.

Section 7 concludes the survey.

Appendix A focuses on listing and providing links to Arabic resources that can be useful for IR such as test collections, stemmers, index tools, and translation tools.

Bibliography

[1] Abdul-Monem Abdul-Al-Aal. 1987. An-nahw ashamil. Maktabat Annahda Al-Masriya, Cairo, Egypt, 1987.

[2] Nasreen AbdulJaleel, Leah S. Larkey. 2003. Statistical transliteration for English-Arabic cross-language information retrieval. In Proceedings of the 2003 Conference on Information and Knowledge Management (CIKM), New Orleans, Louisiana, USA.

[3] Abdelrahim Abdelsapor, Noha Adly, Kareem Darwish, Ossama Emam, Walid Magdy, and Magdy Nagi. 2006. Building a heterogeneous information retrieval collection of printed Arabic documents. In Proceedings of the 2006 Language Resources and Evaluation Conference.

[4] Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2008. Improving Q/A using Arabic Wordnet. In Proceedings The 2008 International Arab Conference on Information Technology (ACIT'2008), Tunisia.

[5] Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2009. Three-level approach for passage retrieval in Arabic question/answering systems. In Proceedings Of the 3rd International Conference on Arabic Language Processing CITALA2009, Rabat, Morocco.

[6] Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2012. IDRAAQ: new Arabic question answering system based on query expansion and passage retrieval." In Proceedings of the 2012 Conference and Labs of the Evaluation Forum (CLEF) (Online Working Notes/Labs/Workshop).

[7] Hani Abu-Salem, Mahmoud Al-Omari, and Martha Evens. 1999. Stemming methodologies over individual query words for Arabic information retrieval. Journal of the American Society for Information Science and Technology Vol. 50(6): p.524-529.

[8] Farooq Ahmad and Grzegorz Kondrak. 2005. Learning a spelling error model from search query logs. In Proceedings of the 2005 Human Language Technology (HLT).

[9] Mohamed Attia Ahmed. 2000. A large-scale computational processor of the Arabic morphology, and applications. Masters thesis in Faculty of Engineering, Cairo University, Cairo, Egypt.

[10] Eneko Agirre, Koldo Gojenola, Kepa Sarasola, and A. Voutilainen. 1998. Towards a single proposal in spelling correction. In the Proceedings of the 1998 International Conference on Computational Linguistics: The Association for Computational Linguistics (COLING ACL '98). San Francisco, CA, pp. 22-28.

[11] Mohammed Aljlayl, Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David O. Holmes, M. Lee, David A. Grossman, and Ophir Frieder. 2001. IIT at TREC-10. In 2001 Text REtrieval Conference (TREC). Gaithersburg, MD.

[12] Mohammed Aljlayl and Ophir Frieder. 2002. On Arabic search: improving the retrieval effectiveness via a light stemming approach. In Proceedings of 2002 Conference on Information and Knowledge Management (CIKM).

[13] Ibrahim Al-Kharashi and Martha Evens. 1994. Comparing words, stems, and roots as index terms in an Arabic information retrieval system. Journal of the American Society for Information Science and Technology, 45(8): pp. 548-560.

[14] May Allam. 1995. Segmentation versus segmentation-free for recognizing Arabic text. In IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology, pp. 228-235.

[15] Imad Al-Sughaiyer and Ibrahim A. Al-Kharashi. 2004. Arabic morphological analysis techniques: A comprehensive survey. In Journal of the American Society for Information Science and Technology Archive, Vol. 55 Issue 3.

[16] Kenneth Beesley. 1996. Arabic finite-state morphological analysis and generation. In the Proceedings of the 1996 International Conference on Computational Linguistics (COLING).

[17] Kenneth Beesley, Tim Buckwalter, and Stuart Newton. 1989. Two-level finite-state analysis of Arabic morphology. In the Seminar on Bilingual Computing in Arabic and English, Cambridge, England.

[18] Yassine Benajiba and Paolo Rosso. 2007. Arabic question answering. Diploma of advanced studies. Technical University of Valencia, Spain.

[19] Yassine Benajiba, Mona Diab, and Paolo Rosso. 2008. Arabic named entity recognition using optimized feature sets. In the Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 284âĂŞ293.

[20] Romaric Besançon, Stéphane Chaudiron, Djamel Mostefa, Olivier Hamon, Ismaïl Timimi, and Khalid Choukri. 2009. Overview of CLEF 2008 INFILE Pilot Track. Evaluating Systems for Multilingual and Multimodal Information Access, in the 2009 Cross-Language Evaluation Forum (CLEF), pp. 939-946.

[21] Romaric Besançon, Stéphane Chaudiron, Djamel Mostefa, Olivier Hamon, Ismaïl Timimi, and Khalid Choukri. 2010. Information filtering evaluation: Overview of CLEF 2009 INFILE Track. In 2010 Conference on Multilingual and Multimodal Information Access Evaluation (CLEF-2010). Text Retrieval Experiments. Springer Berlin Heidelberg, 342-353.

[22] Pinaki Bhaskar, Partha Pakray, Somnath Banerjee, Samadrita Banerjee, Sivaji Bandyopadhyay, and Alexander F. Gelbukh. 2012. Overview of QA4MRE at CLEF 2012: Question answering for machine reading evaluation. In 2012 Conference and Labs of the Evaluation Forum (CLEF) (Online Working Notes/Labs/Workshop).

[23] Fadi Biadsy, Nizar Habash, and Julia Hirschberg. 2009. Improving the Arabic pronunciation dictionary for phone and word recognition with linguistically-based pronunciation rules. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL (HLT NAACL '09), pp. 397-405.

[24] Fadi Biadsy. 2011. Automatic dialect and accent recognition and its application to speech recognition. Ph.D. Thesis, Columbia University.

[25] Fadi Biadsy, Pedro J. Moreno, and Martin Jansche. 2012. Google's cross dialect Arabic voice search. In the Proceedings of the 2012 International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[26] William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Introducing the Arabic WordNet Project. In Proceedings of the third International WordNet Conference (GWC-06).

[27] Mohamed Bouguessa, Benoît Dumoulin, and Shengrui Wang. 2008. Identifying authoritative actors in question-answering forums: The case of Yahoo! answers. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). pp. 866-874.

[28] Eric Brill and Robert Moore. 2000. An improved error model for noisy channel spelling correction. In the Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL), Hong Kong, pp. 286-293.

[29] Chris Buckley and Ellen M. Voorhees. 2004. Retrieval evaluation with incomplete information. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.

[30] Robert Burgin. 1992. Variations in relevance judgments and the evaluation of retrieval performance. Information Processing & Management, Vol. 28(5): pp. 619-627.

[31] Tim Buckwalter. 2002. Buckwalter Arabic morphological analyzer version 1.0. Technical Report LDC2002L49, Linguistic Data Consortium.

[32] Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. ClueWeb09 data set. `http://lemurproject.org/clueweb09/`.

[33] Huaigu Cao, Rohit Prasad, and Prem Natarajan. 2011. Handwritten and typewritten text identification and recognition using hidden Markov models. In 2011 International Conference on Document Analysis and Recognition (ICDAR), pp. 744-748.

[34] Ben Carterette, James Allan, and Ramesh Sitaraman. 2006. Minimal test collections for retrieval evaluation. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 268-275.

[35] Jim Chan, Celal Ziftci, and David Forsyth. 2006. Searching off-line Arabic documents. In the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2.

[36] Aitao Chen and Fredric Gey. 2002. Building an Arabic stemmer for information retrieval. In 2002 Text REtrieval Conference (TREC), Gaithersburg, MD.

[37] Jinying Chen, Rohit Prasad, Huaigu Cao, and Premkumar Natarajan. 2013. Detecting OOV names in Arabic handwritten data. In the 12th International Conference on Document Analysis and Recognition.

[38] David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic dialects. In Proceedings of the European Chapter of ACL (EACL), Vol. 111, pp. 112.

[39] Kenneth Church and William Gale. 1991. Probability scoring for spelling correction. Statistics and Computing, 1, pp. 93-103.
Christopher Cieri, David Graff, Mark Liberman, Nii Martey and Stephanie Strassel. 2000. Large, multilingual, broadcast news corpora for cooperative research in topic detection and tracking: The TDT-2 and TDT-3 corpus efforts. In Proceedings of the 2000 Language Resources and Evaluation Conference.

[40] Cyril Cleverdon. 1997. The Cranfield tests on index language devices. In: Spärck-Jones, Karen; Willett, Peter (Eds.): Readings in Information Retrieval. pp. 47-59.

[41] Paul Clough, Henning Müller, and Mark Sanderson. 2005. The CLEF 2004 cross-language image retrieval track. Multilingual Information Access for Text, Speech and Images, Cross-Language Evaluation Forum, pp. 597-613.

[42] Paul Clough, Azzah Al-Maskari, and Kareem Darwish. 2007. Providing multilingual access to Flickr for Arabic users. Evaluation of Multilingual and Multi-modal Information Retrieval. Springer Berlin Heidelberg, 205-216.

[43] Kareem Darwish. 2002. Building a shallow morphological analyzer in one day. In Proceedings of the ACL-2002 Workshop on Computational Approaches to Semitic Languages.

[44] Kareem Darwish. 2003. Probabilistic methods for searching OCR-degraded Arabic text. Ph.D. Thesis, Electrical and Computer Engineering Department, University of Maryland, College Park.

[45] Kareem Darwish. 2013. Arabizi detection and conversion to Arabic. CoRR abs/1306.6755

[46] Kareem Darwish and Ahmed Abdelali. 2014. Using Stem-Templates to improve Arabic POS and Gender/Number Tagging. In International Conference on Language Resources and Evaluation (LREC-2014).

[47] Kareem Darwish and Ahmed Ali. 2012. Arabic retrieval revisited: Morphological hole filling. In the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL).

[48] Kareem Darwish and Ossama Emam. 2005. The effect of blind relevance feedback on a new Arabic OCR degraded text collection. International Conference on Machine Intelligence: Special Session on Arabic Document Image Analysis.

[49] Kareem Darwish, Hany Hassan, and Ossama Emam. 2005. Examining the effect of improved context sensitive morphology on Arabic information retrieval. In Proceedings of the ACL-2005 Workshop on Computational Approaches to Semitic Languages, pp. 25-30.

[50] Kareem Darwish and Douglas W. Oard. 2002. Term selection for searching printed Arabic. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, pp. 261-268.

[51] Kareem Darwish and Douglas W. Oard. 2002. CLIR experiments at Maryland for TREC 2002: Evidence combination for Arabic-English retrieval. In 2002 Text REtrieval Conference (TREC), Gaithersburg, MD.

[52] Kareem Darwish and Douglas W. Oard. 2003. Balanced query methods for improving OCR-based retrieval. Proceedings 2003 Symposium on Document Image Understanding Technology.

[53] Kareem Darwish and Walid Magdy. 2007. Error correction vs. query garbling for Arabic OCR document retrieval. In ACM Transactions on Information Systems (TOIS), Vol. 26.

[54] Kareem Darwish. 2013. Named Entity Recognition using Cross-lingual Resources: Arabic as an Example. In the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1558-1567.

[55] Ludovic Denoyer and Patrick Gallinari. 2006. The Wikipedia XML corpus. ACM Special Interest Group on Information Retrieval Forum, 40 (1).

[56] Kareem Darwish, Walid Magdy, and Ahmed Mourad. 2012. Language processing for Arabic microblog retrieval. In Proceedings of Conference on Information and Knowledge Management (CIKM).

[57] Anne De Roeck and Waleed El-Fares. 2000. A morphologically sensitive clustering algorithm for identifying Arabic roots. In the Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL), Hong Kong, pp. 199-206.

[58] Mona Diab. 2009. Second generation tools (AMIRA 2.0): Fast and robust tokenization, POS tagging, and base phrase chunking. 2nd International Conference on Arabic Language Resources and Tools.

[59] David Doermann. 1998. The indexing and retrieval of document images: A survey. Computer Vision and Image Understanding, 70(3): pp. 287-298.

[60] Youssef El-Dakar, Khalid El-Gazzar, Noha Adly, Magdy Nagi. 2005. The Million Book Project at Bibliotheca Alexandrina. Journal of Zhejiang University-Science A 6, no. 11 (2005): 1327-1340.

[61] Ali El Kahki, Kareem Darwish, Ahmed Saad El Din, and Mohamed Abd El-Wahab. 2012. Transliteration mining using large training and test sets. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12). pp. 243-252.

[62] Ali El Kahki, Kareem Darwish, Ahmed Saad El Din, Mohamed Abd El-Wahab, Ahmed Hefny, and Waleed Ammar. 2011. Improved transliteration mining using graph reinforcement. In the 2011 Empirical Methods in Natural Language Processing (EMNLP). pp. 1384-1393.

[63] Ahmed El-Kholy and Nizar Habash. 2010. Techniques for Arabic morphological detokenization and orthographic denormalization. In Proceedings of the 2000 Language Resources and Evaluation Conference.

[64] Alexander Fraser, Jinxi Xu, and Ralph M. Weischedel. 2002. TREC 2002 cross-lingual retrieval at BBN. In 2002 Text REtrieval Conference (TREC), Gaithersburg, MD.

[65] Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Jian Hu, Kam-Fai Wong, and Hsiao-Wuen Hon. 2007. Cross-lingual query suggestion using query logs of different languages. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 463-470.

[66] Wei Gao, Cheng Niu, Ming Zhou, and Kam-Fai Wong. 2009. Joint ranking for multilingual web search. In the 2009 European Conference on Information Retrieval (ECIR), pp. 114-125.

[67] Wei Gao, Cheng Niu, Jian-Yum Nie, Ming Zhou, Kam-Fai Wong, and Hsiao-Wuen Hon. 2010. Exploiting query logs for cross-lingual query suggestions. ACM Transactions on Information Systems (TOIS), Vol. 28(2), 6.

[68] Fredric Gey and Douglas W. Oard. 2001. The TREC-2001 Cross- Language Information Retrieval Track: Searching Arabic using English, French or Arabic queries. In 2001 Text REtrieval Conference (TREC), Gaithersburg, MD, pp. 16-23.

[69] Andrew Gillies, Erik Erlandson, John Trenkle, and Steve Schlosser. 1997. Arabic text recognition system. The Symposium on Document Image Understanding Technology.

[70] Julio Gonzalo, Jussi Karlgren, and Paul Clough. 2007. iCLEF 2006 overview: Searching the Flickr WWW photo-sharing repository. Evaluation of Multilingual and Multi-modal Information Retrieval, In 2007 Cross-Language Evaluation Forum (CLEF).

[71] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 267-274.

[72] Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In Proceedings of the Conference of American Association for Computational Linguistics.

[73] Nizar Habash and Owen Rambow. "MAGEAD: a morphological analyzer and generator for the Arabic dialects." Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics.

[74] Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers (NAACL-Short '06), pp. 49-52.

[75] Nizar Habash and Owen Rambow. 2007. Arabic Diacritization through Full morphological tagging. In Proceedings of the 2007 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '07), Companion Volume, pp. 53-56, Rochester, NY.

[76] Nizar Habash, Abdelhadi Soudi, and Timothy Buckwalter. 2007. On Arabic transliteration. Arabic Computational Morphology. Text, Speech and Language Technology, Vol. 38, 2007, pp. 15-22.

[77] Nizar Habash, Owen Rambow, and Ryan Roth. 2009. Mada+Tokan: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

[78] Nizar Habash. 2010. Introduction to Arabic language processing. Synthesis Lectures on Human Language Technologies 3(1), pp. 1-187.

[79] Nizar Habash, Mona T. Diab, and Owen Rambow. 2012. Conventional orthography for dialectal Arabic. In Proceedings of the 2012 Language Resources and Evaluation Conference.

[80] Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological analysis and disambiguation for dialectal Arabic. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '13), pp. 426-432.

[81] Bassam Hammo, Hani Abu-Salem, and Steven Lytinen. 2002. QARAB: a question answering system to support the Arabic language. In Proceedings of the ACL-2002 Workshop on Computational Approaches to Semitic Languages.

[82] Stephen M. Harding, W. Bruce Croft, and C. Weir. Probabilistic retrieval of OCR-degraded text using n-grams. European Conference on Digital Libraries.

[83] Khosrow M. Hassibi. 1994a. Machine printed Arabic OCR. The 22nd AIPR Workshop: Interdisciplinary Computer Vision, SPIE Proceedings.

[84] Khosrow M. Hassibi. 1994b. Machine printed Arabic OCR using neural networks. The 4th International Conference on Multi-lingual Computing, London.

[85] David Hawking. 1996. Document retrieval in OCR-scanned text. 6th Parallel Computing Workshop, Kawasaki, Japan.

[86] Daqing He, Douglas W. Oard, Jianqiang Wang, Jun Luo, Dina Demner-Fushman, Kareem Darwish, Philip Resnik, Sanjeev Khudanpur, Michael Nossal, Michael Subotin, and Anton Leuski. 2003. Making MIRACLEs: Interactive translingual search for Cebuano and Hindi. ACM Transactions on Asian Language Information Processing (TALIP) Vol. 2 Issue 3.

[87] Ahmed Hefny, Kareem Darwish, and Ali Alkahky. 2011. Is a query worth translating: Ask the users! In the 2011 European Conference on Information Retrieval (ECIR), pp. 238-250.

[88] Ismail Hmeidi, Ghassan Kanaan, and Martha Evens. 1997. Design and implementation of automatic indexing for information retrieval with Arabic documents. Journal of the American Society for Information Science and Technology Vol. 48(10): pp. 867-881.

[89] Tao Hong. 1995. Degraded text recognition using visual and linguistic context. Ph.D. thesis, Computer Science Department, SUNY Buffalo, Buffalo, NY.

[90] Dan Jurafsky and James Martin. 2000. Speech and language processing. Prentice Hall.

[91] Paul Kantor and Ellen Voorhees. 1996. Report on the TREC-5 Confusion Track. In 1996 Text REtrieval Conference (TREC), Gaithersburg, MD.

[92] Tapas Kanungo, Gregory Marton, and Osama Bulbul. 1999. OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products. in SPIE Conference on Document Recognition and Retrieval (VI), San Jose, California.

[93] Tapas Kanungo, Osama Bulbul, Gregory Marton, and Doe-Wan Kim. 1997. Arabic OCR systems: State of the art. Symposium on Document Image Understanding Technology, Annapolis, MD.

[94] Shereen Khoja and Roger Garside. 2001. Automatic tagging of an Arabic corpus using APT. The Arabic Linguistic Symposium (ALS), University of Utah, Salt Lake City, Utah.

[95] George Kiraz. 1998. Arabic computation morphology in the West. in The 6th International Conference and Exhibition on Multi-lingual Computing.

[96] Kazuaki Kishida. 2008. Prediction of performance of cross-language information retrieval using automatic evaluation of translation. Library & Info. Science Research. Vol. 30 (2), pp. 138-144.

[97] Wessel Kraaij, Paul Over, and A. Smeaton. 2006. TRECVID 2006-an introduction. In 2006 TREC Video Retrieval Evaluation.

[98] Lori Lamel, Abdelkhalek Messaoudi, and Jean-Luc Gauvain. 2007. Improved acoustic modeling for transcribing Arabic broadcast data. Interspeech'07, pp. 2077-2080.

[99] Leah S. Larkey, James Allan, Margaret E. Connell, Alvaro Bolivar, and Courtney Wade. 2002. UMass at TREC 2002: Cross language and novelty tracks. In 2002 Text REtrieval Conference (TREC).

[100] Leah S. Larkey, Lisa Ballesteros, and Margaret E. Connell. 2002. Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275-282. 77777

[101] Leah S. Larkey, Nasreen AbdulJaleel, and Margaret Connell. 2003. What's in a name?: Proper names in Arabic cross-language information retrieval. Technical report, CIIR Technical Report, IR-278.

[102] Leah S. Larkey, Fangfang Feng, Margaret Connell, and Victor Lavrenko. 2004. Language-specific models in multilingual topic tracking. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 402-409.

[103] Chia-Jung Lee, Chin-Hui Chen, Shao-Hang Kao, and Pu-Jen Cheng. 2010. To translate or not to translate? In Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[104] Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam, and Hany Hassan. 2003. Language model based Arabic word segmentation. In the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, July 2003, Sapporo, Japan. pp. 399-406.

[105] Gina Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-based techniques for cross-language information retrieval. Information Processing & Management Journal, Vol. 41 Issue 3.

[106] Dirk Lewandowski. 2012. Web search engine research. Series editor Amanda Spink. Vol. 4. Emerald Group Publishing.

[107] Wen-Cheng Lin and Hsin-Hsi Chen. 2003. Merging mechanisms in multilingual information retrieval. Lecture notes in computer science, pp. 175-186.

[108] Zhidong A. Lu, Issam Bazzi, Andras Kornai, John Makhoul, Premkumar S. Natarajan, and Richard Schwartz. 1999. A robust, language-independent OCR system. in The 27th AIPR Workshop: Advances in Computer Assisted Recognition, SPIE.

[109] Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic treebank: Building a large-scale annotated Arabic corpus. In NEMLAR Conference on Arabic Language Resources and Tools, pp. 102-109.

[110] Mohamed Maamouri, Dave Graff, Basma Bouziri, Sondos Krouna, Ann Bies, and Seth Kulick. 2010. LDC Standard Arabic Morphological Analyzer (SAMA) version 3.1. Linguistics Data Consortium, Catalog No. LDC2010L01.

[111] Walid Magdy. 2013. TweetMogaz: a news portal of tweets. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1095-1096.

[112] Walid Magdy, Ahmed Ali, and Kareem Darwish. 2012. A summarization tool for time-sensitive social media. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM), pp. 2695-2697.

[113] Walid Magdy and Kareem Darwish. 2006. Arabic OCR error correction using character segment correction, language modeling, and shallow morphology. In the 2006 Empirical Methods in Natural Language Processing (EMNLP). Sydney, Australia, pp. 408-414.

[114] Walid Magdy and Kareem Darwish. 2006. Word-based correction for retrieval of Arabic OCR degraded documents. String Processing and Information Retrieval (SPIRE).

[115] Walid Magdy and Kareem Darwish. 2010. Omni font OCR error correction with effect on retrieval. International Conference on Intelligent Systems Design and Applications (ISDA), 2010, pp. 415-420.

[116] Walid Magdy, Kareem Darwish, Ossama Emam, and Hany Hassan. 2007. Arabic cross-document person name normalization. In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, pp. 25-32.

[117] Walid Magdy, Kareem Darwish, and Mohsen Rashwan. 2007. Fusion of multiple corrupted transmissions and its effect on information retrieval. In the Conference of the Egyptian Society of Language Engineering (ESOLE) 2007.

[118] Walid Magdy and Kareem Darwish. 2008. Effect of OCR error correction on Arabic retrieval. Information Retrieval Journal. 11, 5, 405-425

[119] Walid Magdy, Kareem Darwish, and Motaz El-Saban. 2009. Efficient language-independent retrieval of printed documents without OCR. String Processing and Information Retrieval (SPIRE).

[120] Lidia Mangu, Hong-Kwang Kuo, Stephen Chu, Brian Kingsbury, George Saon, Hagen Soltau, and Fadi Biadsy. 2011. The IBM 2011 GALE Arabic speech transcription system. 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 272-277.

[121] James Mayfield, Paul McNamee, C. Costello, C. Piatko, and A. Banerjee. 2001. JHU/APL at TREC 2001: Experiments in filtering and in Arabic, video, andWeb retrieval. In 2001 Text REtrieval Conference (TREC), Gaithersburg, MD.

[122] J. Scott McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval? Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics.

[123] Paul McNamee and James Mayfield. 2002. Comparing cross-language query expansion techniques by degrading translation resources. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[124] Paul McNamee, C. Piatko, James Mayfield. 2002. JHU/APL at TREC 2002: Experiments in filtering and Arabic retrieval. In 2002 Text REtrieval Conference (TREC).

[125] Mohammed Moussa, Mohamed Waleed Fakhr, and Kareem Darwish. 2012. Statistical denormalization for Arabic Text. In Empirical Methods in Natural Language Processing, pp. 228. 2012.

[126] ThuyLinh Nguyen, Stephan Vogel, and Noah A. Smith. 2010. Nonparametric word segmentation for machine translation. In Proceedings of the 2010 International Conference on Computational Linguistics (COLING), pp. 815-823.

[127] Stefanie Nowak, and Peter Dunker. 2010. Overview of the CLEF 2009 Large Scale Visual Concept Detection and Annotation Task. Multilingual Information Access Evaluation II. Multimedia Experiments Lecture Notes in Computer Science Vol. 6242, pp. 94-109.

[128] Douglas W. Oard, Bonnie Dorr. 1996. A survey of multilingual text retrieval. UMIACS, University of Maryland, College Park.

[129] Douglas W. Oard and Fredric Gey. 2002. The TREC 2002 Arabic/English CLIR Track. In 2002 Text REtrieval Conference (TREC), Gaithersburg, MD.

[130] Douglas W. Oard and William Webber. 2013. Information retrieval for e-discovery. Foundations and Trends in Information Retrieval, Vol. 7, No 1 (2013) 1-145.

[131] Kemal Oflazer. 1996. Error-tolerant finite state recognition with applications to morphological analysis and spelling correction. Computational Linguistics, 22(1), 73-89.

[132] Joseph Olive, Caitlin Christianson, and John McCary. 2011. Handbook of natural language processing and machine translation. Springer ISBN 978-1-4419-7712-0

[133] Iadh Ounis, Craig Macdonald, Jimmy Lin, and Ian Soboroff. 2011. Overview of the TREC-2011 Microblog Track. In 2011 Text REtrieval Conference (TREC).

[134] Lawrence Page. 1998. Method for node ranking in a linked database. US Patent No. 6285999

[135] Jiaul H. Paik, Dipasree Pal, and Swapan K. Parui. 2011. A novel corpus-based stemming algorithm using co-occurrence statistics. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval.

[136] Arfath Pasha, Mohammad Al-Badrashiny, Mohamed Altantawy, Nizar Habash, Manoj Pooleery, Owen Rambow and Ryan M. Roth. 2013. DIRA: Dialectal Arabic Information Retrieval Assistant. The Companion Volume of the Proceedings of International Joint Conference on Natural Language Processing (IJCNLP) 2013: System Demonstrations, pp. 13-16.

[137] Ari Pirkola. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 55-63.

[138] Stephen Robertson and Karen Spärck Jones. 1996. Simple, proven approaches to text-retrieval. Technical Report 356, Computer Laboratory, University of Cambridge, Cambridge, England.

[139] Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran, Musa Alkhalifa, M. Antonia Martí, William Black, Sabri Elkateb, J. Kirk, Adam Pease, Piek Vossen, and Christiane Fellbaum. 2008. Arabic WordNet: Current state and future extensions. In The Fourth Global WordNet Conference, Szeged, Hungary.

[140] Gregory Tassey, Brent R. Rowe, Dallas W. Wood, Albert N. Link, and Diglio A. Simoni. 2010. Economic impact assessment of NIST's Text REtrieval Conference (TREC) Program. Report prepared for National Institute of Technology (NIST).

[141] Khaled Shaalan and Hafsa Raza. 2007. Person Name Entity Recognition for Arabic. In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, pp. 17âĂŞ24, Prague, Czech Republic.

[142] Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013. Translating dialectal Arabic to English. The 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria.

[143] Shirin Saleem, Huaigu Cao, Krishna Subramanian, Matin Kamali, Rohit Prasad, Prem Natarajan. 2009. Improvements in BBN's HMM-based offline Arabic handwriting recognition system. 10th International Conference on Document Analysis and Recognition.

[144] Wael Salloum and Nizar Habash. 2012. Elissa: A dialectal to standard arabic machine translation system. In Proceedings of the 2012 International Conference on Computational Linguistics (COLING). Mumbai, India.

[145] Gerard Salton and M. Lesk. 1969. Relevance assessments and retrieval system evaluation. Information Storage and Retrieval, 1969 (4), pp. 343-359.

[146] Mark Sanderson. 1994. Word sense disambiguation and information retrieval. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 142-151.

[147] Mark Sanderson and H. Joho. 2004. Forming test collections with no system pooling. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, South Yorkshire, UK.

[148] Mark Sanderson. 2010. Test collection based evaluation of information retrieval systems. Foundations and Trends in Information Retrieval, Vol. 4, No. 4, pp. 247-375.

[149] Jacques Savoy, and Yves Rasolofo. 2002. Report on the TREC 11 experiment: Arabic, named page and topic distillation searches. In 2002 Text REtrieval Conference (TREC).

[150] Asad Sayeed, Tamer Elsayed, Nikesh Garera, David Alexander, Tan Xu, Douglas W. Oard, David Yarowsky, and Christine Piatko. 2009. Arabic cross-document coreference detection. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 357-360.

[151] Mohammed Q. Shatnawi, Qusai Q. Abuein, and Omar Darwish. 2011. Verification hadith correctness in islamic web pages using information retrieval techniques. Proceedings of International Conference on Information & Communication Systems, Irbid, Jordan.

[152] Mohammed Q. Shatnawi, Muneer Bani Yassein, and Reem Mahafza. 2012. A framework for retrieving Arabic documents based on queries written in Arabic slang language. Journal of Information Science, Vol. 38, no. 4: 350-365.

[153] Luo Si and Jamie Callan. 2006. CLEF 2005: Multilingual retrieval by combining multiple multilingual ranked lists. Accessing Multilingual Information Repositories Lecture Notes in Computer Science, Vol. 4022, pp. 121-130.

[154] Amit Singhal, Gerard Salton, and Chris Buckley. 1996. Length normalization in degraded text collections. The 5th Annual Symposium on Document Analysis and Information Retrieval.

[155] Alan F. Smeaton, Paul Over, and Wessel Kraaij. 2006. Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM international workshop on Multimedia information retrieval, pp. 321-330.

[156] Ian Soboroff, Charles Nicholas, and Patrick Cahan. 2001. Ranking retrieval systems without relevance judgments. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[157] Ian Soboroff and Stephen Robertson. 2003. Building a filtering test collection for TREC 2002. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 243-250.

[158] Hagen Soltau, George Saon, Brian Kingsbury, Jeff Kuo, Lidia Mangu, Daniel Povey, and Geoffrey Zweig. 2007. The IBM 2006 GALE Arabic ASR system. In Proceedings of the 2007 International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 4, pp. 349-352.

[159] Hagen Soltau, Lidia Mangu, and Fadi Biadsy. 2011. From Modern Standard Arabic to Levantine ASR: Leveraging GALE for dialects. In Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, pp. 266-271.

[160] Stephanie Strassel, Mark A. Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda. 2008. Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In Proceedings of the 2008 Language Resources and Evaluation Conference.

[161] Stephanie Strassel. 2009. Linguistic resources for Arabic handwriting recognition. In Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt.

[162] Trevor Strohman, Donald Metzler, Howard Turtle, and W. Bruce Croft. 2005. Indri: A language model-based search engine for complex queries. In Proceedings of the International Conference on Intelligent Analysis, Vol. 2, No. 6, pp. 2-6.

[163] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online QA collections. In Proceedings of the 46th Annual Meeting for the Association for Computational Linguistics: Human Language Technologies (ACL HLT '08), pp. 719-727.

[164] Kazem Taghva, Julie Borasack, Allen Condit, and Jeff Gilbreth. 1994. Results and implications of the noisy data projects. 1994, Information Science Research Institute, University of Nevada, Las Vegas.

[165] Kazem Taghva, Julie Borasack, Allen Condit, and Padma Inaparthy. 1995. Querying short OCR'd documents. 1995, Information Science Research Institute, University of Nevada, Las Vegas.

[166] Kazem Taghva, Julie Borasack, and Allen Condit. 1994. An expert system for automatically correcting OCR output. In SPIE-Document Recognition.

[167] Mikael Tillenius. 1996. Efficient generation and ranking of spelling error corrections. NADA tech. report TRITA-NA-E9621.

[168] Omar Trigui, Lamia Hadrich Belguith, Paolo Rosso, Hichem Ben Amor, and Bilel Gafsaoui. 2012. Arabic QA4MRE at CLEF 2012: Arabic Question Answering for Machine Reading Evaluation. CLEF (Online Working Notes/Labs/Workshop).

[169] Ming-Feng Tsai, Yu-Ting Wang, and Hsin-Hsi Chen. 2008. A study of learning a merge model for multilingual information retrieval. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[170] Yuen-Hsien Tseng and Douglas W. Oard. 2001. Document image retrieval techniques for Chinese. In Symposium on Document Image Understanding Technology (SDIUT). Columbia, MD, pp. 151-158.

[171] Theodora Tsikrika, and Jana Kludas. 2009. Overview of the WikipediaMM Task at ImageCLEF 2009. Multilingual Information Access Evaluation II. Multimedia Experiments, Lecture Notes in Computer Science Vol. 6242, pp 60-71.

[172] Raghavendra Udupa, K. Saravanan, A. Bakalov, and A. Bhole. 2009. "They are out there, if you know where to look": Mining transliterations of OOV query terms for cross-language information retrieval. In the 2009 European Conference on Information Retrieval (ECIR), LNCS 5478, pp. 437-448.

[173] Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. 2009. Mint: A method for effective and scalable mining of named entity transliterations from large comparable corpora. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 799-807.

[174] Ellen Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia.

[175] Jianqiang Wang and Douglas W. Oard. 2006. Combining bidirectional translation and synonymy for cross-language information retrieval. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 202-209.

[176] Charles L. Wayne. 1998. Detection & tracking: A case study in corpus creation & evaluation methodologies. Language Resources and Evaluation Conference, Granada, Spain.

[177] Dan Wu, Daqing He, Heng Ji, and Ralph Grishman. 2008. A study of using an out-of-box commercial MT system for query translation in CLIR. Workshop on Improving non-English web searching, Proceedings of 2008 Conference on Information and Knowledge Management (CIKM).

[178] Jinxi Xu, Alexander Fraser, and Ralph Weischedel. 2002. Empirical studies in strategies for Arabic retrieval. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.

[179] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. 2008. A simple and efficient sampling method for estimating AP and NDCG. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.