

Kareem Darwish
aiXplain Inc., Los Gatos, CA, US

Mobile Phone: +905367218818

E-mail: Kareem.Darwish@aiXplain.com

Google Scholar: <https://scholar.google.com/citations?user=y7tIR6UAAAAJ>

LinkedIn: <https://www.linkedin.com/in/kareem-darwish-788b9125/>

Education

Ph.D. in Electrical and Computer Engineering,
University of Maryland, College Park, MD, May 2003.

Ph.D. Thesis: Probabilistic Methods for Searching OCR-Degraded Arabic Text

GPA: 3.73

Masters Degree in Electrical and Computer Engineering,
University of Maryland, College Park, August 1999.

GPA: 3.40

Bachelor of Electrical and Computer Engineering,
University of Maryland, College Park, December 1995.
Magna Cum Laude, GPA: 3.88

Research and Employment Experience

Principal Scientist, aiXplain Inc., Los Gatos, CA, US

November 2021 – Present

- Leading (as PI) major project with SDAIA, KSA for developing a variety of language and speech technologies including:
 - Pre-training and fine-tuning Arabic Large Language Models (LLMs)
 - Preparing training data for LLM applications including writing guidelines, designing annotation procedures, building quality checks
 - Building Arabic text-to-speech
 - Training Arabic processing models such word segmentation, denormalization, and diacritization
- Leading a major annotation project with SDAIA involving the transcription and labeling of 6,000 hours of audio including:
 - Designing guidelines and procedures
 - Developing pre-annotation models and speed up processing including quality estimation, gender identification, etc.
 - Implementing a rigorous QA process to ensure annotation quality.
- Managing multiple speech processing projects for building:
 - Voice activity detection and speaker diarization
 - Speaker identification
 - Speech classification

Acting Research Director & Principal Scientist, Qatar Computing Research Institute, HBKU, Doha, Qatar

February 2011 – November 2021

Arabic Language Technologies Group

- Defining the priorities of natural language processing research at QCRI
- Leading and participating in collaborations between QCRI and external entities such as Aljazeera.net and Boeing
- Conducting productive research in the following:
 - Farasa Arabic NLP Toolkit
 - Leading the Arabic NLP group
 - Developing algorithms for segmentation, POS tagging, diacritization, named entity recognition, etc.

- Expanding the toolkit to handle standard Arabic and dialects
 - Programming major parts of the toolkit
- Natiq Arabic text-to-speech:
 - Procuring data for TTS including data selection, preparation, and recording
 - Pre-processing text for TTS including custom diacritization and number conversion.
- Asad Arabic text processing toolkit:
 - Building a variety of Arabic datasets and models for emotion detection, sentiment analysis, offensive language detection, hate speech detection, etc.
- Social search
 - Improving language handling for search using language dependent and language independent methods
 - Developing cross media (social media and news) summarization (ex. tweetMogaz.com)
- Social computing
 - Developing technologies for the automated automatic stance detection, polarization quantification, and propaganda accounts detection.
 - Applying the methodologies to specific case studies including turmoil in Egypt, ISIS sympathizers, Islamophobia, and xenophobia.

Researcher, Microsoft, Cairo,

March 2007 – February 2011

Cairo Microsoft Innovation Center (CMIC)

- Conducting productive research in the following projects:
 - BookWeb:
 - Exploring digitized content by cross linking topical segments and automatically extracted key phrases to topical segments
 - Automatically identifying table of content pages and linking their entries to appropriate pages
 - Automatic constructing of tables of content by identifying headlines in digitized content
 - Improving search of digitized content by identifying most valuable parts of content
 - Resulting in: ThinkWeek paper; invention disclosure; collaboration with MSRC (Natasa Milic-Frayling's group); joint TechFest 2008 demo with MSRC; papers in INEX and CIKM workshop.
 - IBIS:
 - Measuring Bing's (formerly Live) search effectiveness for Arabic
 - Working on improving search effectiveness:
 - Index coverage: supplied a white list of good Arabic URL's for manual boosting of their static rank
 - Word breaking: designed, tested, and helped code a new Arabic word breaker
 - Resulting in: a ThinkWeek paper; a white list that helped quadruple Bing's Arabic index; a word-breaker that was checked-in into Bing's tree
 - Transbulletization:
 - Attempting to trim superfluous parts from sentences that are translated from different languages into English using parsing and statistical summarization techniques without breaking the flow of sentences
 - Integrating concept into a cross-language search application
 - Resulting in: TechFest 2009 demo; invention disclosure
 - Enterprise documenting linking:
 - Attempting to help users navigate intranets by providing contextual links/recommendations based on a user's:
 - Current context -- searching using salient term in context
 - Browsing history -- utilizing relevance feedback and filtering
 - Providing related resources such as documentations, people pages, etc.
 - Resulting in: a TechFest 2009 demo; tool to become TechFest 2010 official search/browsing tool; incubation with FuturePoint group
 - Search results diversification:

- Attempting to identify a query's different meanings (extrinsic diversity) or different facets (intrinsic diversity) to diversify top search results
 - Using knowledge bases to perform diversification
 - Using density-based clustering to perform diversification
 - Resulting in: paper submitted to ECIR-2010; participation in TREC-2009 relevance feedback track (in collaboration with MSRC's Stephen Robertson's group) with an oral presentation at TREC; collaboration with PAMI group in University of Waterloo (Dr. Mohamed Kamel)
- OCRless retrieval:
 - Language independent searching of document images without performing OCR by clustering similar connected components and rendering queries into images
 - Applying IR techniques such as weighted structured queries to improve retrieval effectiveness
 - Applying density based clustering to improve clustering
 - Resulting in: TechFest 2009 demo; invention disclosure; SPIRE-2009 paper
- Machine translation:
 - Developing Arabic language handling to improve Ar \leftrightarrow En MT
 - Designing and testing an Arabic word breaker for Ar \Rightarrow En MT to improve lexical coverage
 - Helping design and test Arabic reverse word breaker for En \Rightarrow Ar MT
 - Performing acronym expansion to aid En \Rightarrow any_language MT
 - Detecting named entities as targets for transliteration
 - Transliterating Ar \leftrightarrow En named entities
 - Resulting in: Tech transfer of Arabic word breaker in MSR MT system; reverse word breaker tech transfer expected in Nov. 2009
- Bing instant answers:
 - Performing cross language search to improve image search for Arabic by overcoming the lack of Arabic meta-information about images
 - Arabic queries translated into English and results are shown to user
 - Translating English instant answers automatically into Arabic
 - Translating English Wikipedia Info-boxes into Arabic to aid in on Arabic question answering based on knowledge bases
 - Resulting in: 2 ThinkWeek papers; development of instant answers in progress

Lecturer, Cairo University, Cairo,

August 2005 – May 2016

Information Systems Department, Faculty of Computer Science and Informatics

- Designing and teaching of Courses
 - Data structures
 - Unstructured document retrieval
 - Managing teaching assistants
 - Supervising research assistants in
 - Rapid development of IR test collection
 - Arabic-Hebrew cross language retrieval
 - Employing Arabic morphological analysis in word clustering and highlighting of search results
 - Interactive query expansion
 - Automatic web page structural analysis
 - Wikipedia named entity tagging
 - Supervising senior graduation projects: natural language question authoring; multi-lingual desktop search; affect resolution; browser history caching and search; Arabic named entity recognition
- Egyptian Ministry of Communication and Information Technology Research Center of Excellent (co-PI)
- Developing a web portal for aggregating Arabic web news (www.alzoa.com)
 - Deploying state-of-the-art Arabic text search
 - Using automatic document clustering
 - Investigating automatic Arabic phrase extraction for summarization purposes

- Investigating ways to anonymously customize web pages to suite specific users
- Exploring ways to make news interactive and solicit user input

Research Consultant, IBM Technology Development Center, Cairo,

March 2005 – March 2007

Human Language Technologies Group

- Performing productive research and development in the following areas:
 - Arabic OCR degraded text retrieval
 - Incorporating character, word, stem, and stem-template based language modeling to improve error correction
 - Adapting blind relevance to OCR degraded documents
 - Investigating/developing techniques for developing relevance judgments without pooling for OCR degraded collections
 - Developing degraded Arabic word clustering to improve search term highlighting
 - Machine/Human Assisted Human/Machine translation
 - Using the output of a machine translation system to build localized language models to improve type-ahead for translators
 - Incorporating speech recognition technology with existing machine translation technology to speed-up human translation
 - Information extraction from biomedical text
 - Employing unsupervised learning techniques to infer patterns that contain relationships/interactions between biomedical named entities
 - Applying inferred models in extracting protein-protein interaction
 - Adaptive cross-language text filtering
- Team leader, InfoMind Project
 - Integrating MT, IE, IR, adaptive filtering, and information visualization
 - Managing research engineers in varying parts of the projects

Lecturer, German University in Cairo,

January 2004 – August 2005

Department of Information Engineering and Technology

- Designing and teaching introductory Electrical Engineering Courses
 - Basic Circuit Theory
 - Electric Circuits Lab
 - Digital Logic Design
- Managing teaching and research assistants
- Performing research in Arabic information retrieval
 - Comparing different word stemming and clustering techniques for Arabic information retrieval
 - Devising new methods for rapid construction of test collections for monolingual and cross-lingual retrieval
- Performing research in Bioinformatics
 - Evaluating the effectiveness of controlled vocabularies in information retrieval
 - Designing automatic methods for assigning controlled vocabulary entries to documents
 - Participating in the 2004 Text REtrieval Conference (TREC) Genomics Track (Ranked 3rd in the Adhoc Retrieval task and 6th in the Triage task)
- Collaborating with the Library of Alexandria in the Million Book Project
 - Introduction and evaluation of information retrieval technology to scanned and OCR'ed Arabic books
 - Exploration and evaluation of error tolerant Arabic word clustering and morphology techniques

Senior Consultant, the KEVRIC Company, Silver Spring, MD

June, 2003 – February 2004

Knowledge Management project – under contract from the National Institutes of Health (NIH)

- Served as Principal Investigator (PI) on the project to evaluate emerging knowledge management technologies intended to facilitate NIH's grant review process

- Identifying the needs of scientific review administrators tasked with routing incoming grant proposals to appropriate reviewers
- Researching existing technologies and developing alternative ones to address the stated needs
- Establishing criteria for evaluating potentially viable technologies
- Evaluating the usability and effectiveness of the different technologies

The BISC project – under contract from NIH

- Served as a research scientist to facilitate the adoption and integration of varying biomedical ontologies intended to support clinical research applications
- Migrating different biomedical ontologies in varying formats (such DAML and OWL) to standardized formats to facilitate their integration
- Designing applications that rely on the ontologies and defining insertion points for integrating the ontologies into the applications
- Working on interfacing applications and the ontologies using API's

Research Assistant, University of Maryland, College Park

April, 2000 – May, 2003

University of Maryland Institute for Advanced Computer Studies,
Computational Linguistics and Information Processing Laboratory (CLIP)

- Developed systems for Arabic OCR-based retrieval
 - Manual and Automatic test collection construction
 - Development and verification of OCR-degradation models
 - Methods of enhancing OCR-based retrieval using OCR-degradation models and structured query methods
- Developed systems for cross-language information retrieval (CLIR):
 - Searching English, Arabic, and French documents using Arabic, English, or French queries
 - Mining translation resources from the web and OCR-degraded dictionaries
 - Combining translation resources
 - Construction of test collections
- Built state of the art Arabic morphology tools for Information Retrieval
 - Designing, implementing, and training Sebawai, a statistical morphology engine
 - Developing Al-Stem, a light Arabic stemmer
- Participated in international CLIR evaluations:
 - TREC 2001 and 2002

Principal and Managing Officer, Al-Areeb Electronic Publishers, Laurel, Maryland

July 1997 – February 1999

- Researched the market for commercially available packages to target software
- Supervised and programmed a full-featured text retrieval system that includes a full-in-text search engine, topic index capabilities, and scalable text display and printing system.
- Programmed Web Based Applications integrating state of the art audio compression schemes and database sockets
- Managed 4 permanent employees including 2 programmers and hired and managed additional temporary employees
- Produced many commercial products and prepared package designs, CD pre-masters, user manuals, and packaging
- Overlooked product marketing (more than 40,000 copies of products sold)
- Wrote many business offers and proposals

Research and Teaching Assistant, University of Maryland College Park

Sept. 1996 - June 1997

Electrical and Computer Engineering Department

- Designed and implemented the graduate students and applicants database for the Electrical and computer Engineering Department using MS Access DBMS and Visual Basic
- Wrote scripts to interface between the database and a UNIX system

- Programmed complex database tasks using VBA.

Scientific Assistant, Kuwait University

February 1996 - August 1996

Electrical Engineering Department

- Performed tasks of a teaching assistant for Software Engineering, Microsoft Access, and Fortran
- Assisted in the administration of the Sun-Sparc Lab

Research Assistant, University of Maryland College Park

July 1995 - December 1995

Electrical Engineering Department

- Helped port an optimization package called X-Console from UNIX to Microsoft Windows
- Wrote portions of the documentation of the package along with generating sample files and detailed examples

Electrical Engineering Intern, Army Research Lab, Adelphi, Maryland

July 1994 - Feb. 1995

Digital Imaging Group

- Designed digital boards using Hardware Description Language (VHDL) using Altera and PowerView
- Implemented Digital Boards using Programmable Gate Arrays and other commercially available digital components. The boards were later used as a part of a DSP system

Publications

Book:

Kareem Darwish and Walid Magdy. Arabic Information Retrieval. In Foundations and Trends® in Information Retrieval 7, 4, February 2014.

Book Chapters:

Hany Hassan, Kareem Darwish. 2014. Statistical Machine Translation. Natural Language Processing of Semitic Languages, pp. 199-219, Springer Berlin Heidelberg

Kareem Darwish. 2014. Information Retrieval. Natural Language Processing of Semitic Languages, pp. 299-334, Springer Berlin Heidelberg

Darwish, K. and D. Oard. Adapting Morphology for Arabic Information Retrieval. Editors: Abdelhadi Souidi, Günter Neumann and Antal Van den Bosch, Arabic Computational Morphology: Knowledge-based and Empirical Methods. In Kluwer/Springer's series on Text, Speech, and Language Technology, 2006.

Darwish, K., M. Diab, and N. Habash, Eds. Computational Approaches to Semitic Languages. Workshop Proceedings. Association for Computational Linguistics, Ann Arbor, Michigan, 2005.

Journal Papers:

Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak, Mohamed Eldesouki. 2021. Arabic diacritic recovery using a feature-rich bilstm model. Transactions on Asian and Low-Resource Language Information Processing, (20) 2: 1-18.

Kareem Darwish, Mohammed Attia, Hamdy Mubarak, Younes Samih, Ahmed Abdelali, Lluís Màrquez, Mohamed Eldesouki, Laura Kallmeyer. 2020. Effective multi-dialectal Arabic POS tagging. Natural Language Engineering, (26) 6: 677-690.

Amr Elalfy, Kareem Darwish, Olaf Weber. 2020. Corporations and Sustainable Development Goals Communication on Social Media: Corporate Social Responsibility or just Another Buzzword? Journal of Sustainable Development.

Kareem Darwish, Mohammed Attia, Hamdy Mubarak, Younes Samih, Ahmed Abdelali, Lluís Màrquez, Mohamed Eldesouki, Laura Kallmeyer. 2020. Effective multi-dialectal Arabic POS tagging. Journal for Natural Language Engineering. 2020.

Kareem Darwish, Walid Magdy, Afshin Rahimi, Tim Baldwin, Norah Abokhodair. Predicting Online Islamophobic Behavior After #ParisAttacks. The Journal of Web Science 3 (1).

- Salvatore Romeo, Giovanni Da San Martino, Yonatan Belinkov, Alberto Barrón-Cedeño, Mohamed Eldesouki, Kareem Darwish, Hamdy Mubarak, James Glass, Alessandro Moschitti. 2017. Language processing and learning models for community question answering in Arabic. *Information Processing & Management*, 2017.
- Magdy, Walid, Kareem Darwish, and Ingmar Weber. #FailedRevolutions: Using Twitter to study the antecedents of ISIS support. *First Monday* 21.2 (2016).
- Magdy, W. and K. Darwish. Effect of OCR error correction on Arabic retrieval. *Journal of Information Retrieval*, Vol. 11, No. 5, 405-425, 2008
- Darwish, K. and W. Magdy. Error correction vs. query garbling for Arabic OCR document retrieval. *ACM Transactions on Information Systems (TOIS)* Vol. 26, Issue 1, Article 5, 2007
- He, D., D. W. Oard, J. Wang, J. Luo, D. Demner-Fushman, K. Darwish, P. Resnik, S. Khudanpur, M. Nossal and A. Leuski. Making MIRACLES: Interactive Translingual Search for Cebuano and Hindi. To appear in *ACM Transactions of Asian Language Information Processing*, 2003.

Refereed Conference Papers:

- Kareem Darwish. 2022. News consumption in time of conflict: 2021 Palestinian-Israel War as an example. 14th ACM Web Science Conference 2022, pp. 260-268.
- Shreyas Sharma, Kareem Darwish, Lucas Pavanelli, Thiago Castro Ferreira, Mohamed Al-Badrashiny, Kamer Ali Yuksel, Hassan Sawaf. 2022. MTLens: Machine Translation Output Debugging. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4221-4226.
- Sabit Hassan, Shaden Shaar, Kareem Darwish. 2022. Cross-Lingual Emotion Detection. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.
- Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, Preslav Nakov. 2021. Fighting the COVID-19 infodemic in social media: a holistic perspective and a call to arms. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, (15) 913-922.
- Ammar Rashed, Mucahid Kutlu, Kareem Darwish, Tamer Elsayed, Cansın Bayrak. 2021. Embeddings-based clustering for target specific stances: The case of a polarized Turkey. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, (15) 537-548.
- Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, Kareem Darwish. 2021. Asad: Arabic social media analytics and understanding. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistic (EACL)*, pp. 113-118.
- Younes Samih, Kareem Darwish. 2021. A few topical tweets are enough for effective user stance detection. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistic (EACL)*, pp. 2637-2646.
- Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the Arab world. *Communications of the ACM*, (64) 4: 72-81.
- Fouzi Harrag, Maria Dabbah, Kareem Darwish, Ahmed Abdelali. 2020. Bert Transformer model for Detecting Arabic GPT2 Auto-Generated Tweets. *Proceedings of the Fifth Arabic Natural Language Processing Workshop*.
- Shammur Absar Chowdhury, Ahmed Abdelali, Kareem Darwish, Jung Soon-Gyo, Joni Salminen, Bernard J Jansen. 2020. Improving Arabic text categorization using transformer training diversification. *Proceedings of the Fifth Arabic Natural Language Processing Workshop*.
- Hamdy Mubarak, Shima Amer, Ahmed Abdelali, Kareem Darwish. 2020. Arabic Curriculum Analysis. *Proceedings of the 28th International Conference on Computational Linguistics*.
- Peter Stefanov, Atanas Atanasov, Kareem Darwish, Preslav Nakov. 2020. Predicting the Topical Stance and Political Leaning of Media using Tweets. In *Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics (ACL 2020)*, Seattle, Washington, USA, July, 2020.
- Kareem Darwish, Michael Aupetit, Peter Stefanov, Preslav Nakov. 2020. Unsupervised User Stance Detection on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM'2020)*. Atlanta, Georgia, USA, June, 2020.
- Hamdy Mubarak, Ahmed Abdelali, Sabit Hassan, Kareem Darwish. 2020. Spam detection on Arabic twitter. *Social Informatics: 12th International Conference, SocInfo 2020*.
- Chereen Shurafa, Kareem Darwish, Wajdi Zaghouni. 2020. Political framing: US COVID19 blame game. *Social Informatics: 12th International Conference, SocInfo 2020*.

- Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, Hend Al-Khalifa. 2020. Overview of OSACT4 Arabic Offensive Language Detection Shared Task. In: OSACT 4: The 4th Workshop on Open-Source Arabic Corpora and Processing Tools (LREC-2020). Marseille, France, May 2020.
- Hamdy Mubarak, Kareem Darwish. 2019. Arabic Offensive Language Classification on Twitter. International Conference on Social Informatics (SocInfo'2019). Doha, Qatar, November 2019.
- Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, Kareem Darwish, James Glass, Preslav Nakov. 2019. Tanbih: Get To Know What You Are Reading. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. Hong Kong, China, November 2019.
- Mohamed Eldesouki, Naassih Gopee, Ahmed Ali, Kareem Darwish. (2019) FarSpeech: Arabic Natural Language Processing for Live Arabic Speech. Proc. Interspeech 2019. Graz, Austria. September 2019.
- Hamdy Mubarak, Ahmed Abdelali, Kareem Darwish, Mohamed Eldesouki, Younes Samih, Hassan Sajjad. 2019. A System for Diacritizing Four Varieties of Arabic. Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'2019): System Demonstrations, Hong Kong, November 2019.
- Hamdy Mubarak, Ahmed Abdelali, Hassan Sajjad, Younes Samih, Kareem Darwish. 2019. Highly effective Arabic Diacritization using Sequence to Sequence Modeling. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'2019), Minneapolis, MN, USA, June 2019.
- Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, Younes Samih, Rehab Alharb, Walid Magdy. 2018. Multi-Dialect Arabic POS Tagging: A CRF Approach. LREC 2018.
- Rehab Alharbi, Walid Magdy, Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak. Part-of-Speech Tagging for Arabic Gulf Dialect Using Bi-LSTM. LREC 2018.
- Kareem Darwish, Dimitar Alexandrov, Preslav Nakov, Yelena Mejova. 2017. Seminar Users in the Arabic Twitter Sphere. International Conference on Social Informatics, pp. 91-108, 2017.
- Kareem Darwish, Walid Magdy, Tahar Zanouda. 2017. Trump vs. Hillary: What Went Viral During the 2016 US Presidential Election. International Conference on Social Informatics, pp. 143-161, 2017.
- Kareem Darwish, Walid Magdy, Tahar Zanouda. 2017. Improved Stance Prediction in a User Similarity Feature Space. 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2017.
- Younes Samih, Mohamed Eldesouki, Mohammed Attia, Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak, Laura Kallmeyer. 2017. Learning from Relatives: Unified Dialectal Arabic Segmentation. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp. 432-441, 2017.
- Abdelali, Ahmed, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 11-16. Association for Computational Linguistics, San Diego, California, 2016.
- Walid Magdy, Kareem Darwish, Afshin Rahimi, Norah Abokhodair and Tim Baldwin. #ISISisNotIslam or #DeportAllMuslims? Predicting Unspoken Views. Web Science (WebSci'16). 2016.
- Kareem Darwish and Hamdy Mubarak. Farasa: Fast and Accurate Arabic Word Segmenter. LREC'16. 2016
- Walid Magdy, Kareem Darwish, Ingmar Weber. 2015. I like ISIS, but I want to watch Chris Nolan's new movie: Exploring ISIS Supporters on Twitter. Proceedings of the 26th ACM Conference on Hypertext & Social Media, pp. 321-322.
- Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and Network Dynamics Behind Egyptian Political Polarization on Twitter. In Proceedings of Computer-Supported Cooperative Work and Social Computing (CSCW). Vancouver, BC, Canada, March 2015.
- Kareem Darwish, Hassan Sajjad, Hamdy Mubarak. Verifiably Effective Arabic Dialect Identification. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), Doha, Qatar, October 2014.
- Kareem Darwish and Wei Gao. Simple Effective Microblog Named Entity Recognition: Arabic as an Example. In Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, May 2014.
- Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak, 2014, Using Stem-Templates to Improve Arabic POS and Gender/Number Tagging. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, May 2014.

- Hassan Sajjad, Kareem Darwish and Yonatan Belinkov. Translating Dialectal Arabic to English. In Proceedings of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria, August 2013
- Kareem Darwish, Walid Magdy, Ahmed Mourad. 2012. Language Processing for Arabic Microblog Retrieval. Proceedings of the 21st ACM international conference on Information and knowledge management, CIKM-2012, pp. 2427-2430.
- Kareem Darwish. 2013. Named Entity Recognition using Cross-lingual Resources: Arabic as an Example. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 1558–1567
- Alok Kothari, Walid Magdy, Kareem Darwish, Ahmed Mourad, Ahmed Taei. 2013. Detecting Comments on News Articles in Microblogs. ICWSM-2013.
- Wei Gao, Peng Li, and Kareem Darwish. Joint Topic Modeling for Event Summarization across News and Social Media Streams. In Proceedings of the 21st ACM Conference on Information and Knowledge Management (CIKM 12), October 29–November 2, 2012, Maui, HI, USA
- Walid Magdy, Ahmed Ali, Kareem Darwish. A summarization tool for time-sensitive social media. Proceedings of the 21st ACM international conference on Information and knowledge management, 2012
- Mohammed Moussa, Mohamed Waleed Fakh, Kareem Darwish. 2012. Statistical Denormalization for Arabic Text. Empirical Methods in Natural Language Processing, Volume: 228.
- Kareem Darwish, Ahmed Ali. “Arabic Retrieval Revisited: Morphological Hole Filling”. ACL-2012, 2012.
- Ali El Kahki, Kareem Darwish, Ahmed Saad El Din, Mohamed Abd El-Wahab. “Transliteration Mining Using Large Training and Test Sets”. NAACL-2012, 2012.
- Ali El-Kahky, Kareem Darwish, Ahmed Saad Aldein, Mohamed Abd El-Wahab, Ahmed Hefny, Waleed Ammar. “Improved Transliteration Mining Using Graph Reinforcement”. EMNLP-2011, 2011.
- Hefny, A., K. Darwish, Ali El-kahki. “Crowdsourcing for Cross Language and Multilingual Web Search”. ECIR-2011, 2011.
- Magdy, W., K. Darwish. “Omni Font OCR Error Correction with Effect on Retrieval” International Conference on Intelligent Systems Design and Applications (ISDA 2010), 2010.
- Darwish, K., A. El Deeb, N. Yousri, M. Kamel. “Omni Font OCR Error Correction with Effect on Retrieval” International Conference on Intelligent Systems Design and Applications (ISDA 2010), 2010.
- Magdy, W., K. Darwish, and M. El-Sabban. Efficient Language-Independent Retrieval of Printed Documents without OCR. Symposium on String Processing and Information Retrieval (SPIRE'09), 334-343, 2009.
- Ibrahim, H., K. Darwish, and A. Abdel-sabor. Automatic extraction of textual elements from news web pages. LREC-2008, 2008.
- Magdy, W., K., Darwish, and M. Rashwan. Fusion of multiple corrupted transmissions and its effect on information retrieval. In Seventh Conference on Language Engineering, ESOLEC, 351–358, 2007.
- Magdy, W. and Darwish, K. Arabic OCR error correction using character segment correction, language modeling, and shallow morphology. Empirical Methods in Natural Language Processing (EMNLP'06). Sydney, Australia, 408-414, 2006.
- Magdy, W. and K. Darwish. Arabic word-based correction for Retrieval of Arabic OCR degraded documents. Symposium on String Processing and Information Retrieval (SPIRE'06). University of Strathclyde, UK, 2006.
- Clough, P., A. Al-Maskari, and K. Darwish. Providing Multilingual Access to Flickr for Arabic Users. Evaluation of Multilingual and Multi-modal Information Retrieval, 205-216, 2006.
- Darwish, K., W. Magdy, O. Emam, A. Abdelsapor, N. Adly, M. Nagi. Building a Heterogeneous Information Retrieval Collection of Arabic Document Images. LREC 2006.
- Darwish, K. and O. Emam. The Effect of Blind Relevance Feedback on a New Arabic OCR Degraded Text Collection. In International Conference on Machine Intelligence: Special Session on Arabic Document Image Analysis, 2005.
- Darwish, K. and D. Oard. Probabilistic Structured Query Methods. In ACM SIGIR: 338-344, 2003.
- Darwish, K. and D. Oard. Term Selection for Searching Printed Arabic. In ACM SIGIR: 261-268, 2002.

Workshop Papers

- Noof Alabbasi, Mohamed Al-Badrashiny, Maryam Aldahmani, Ahmed AIDhanhani, Abdullah Saleh Alhashmi, Fawaghy Ahmed Alhashmi, Khalid Al Hashemi, Rama Emad Alkhobbi, Shamma T Al Maazmi, Mohammed Ali Alyafeai, Mariam M Alzaabi, Mohamed Saqer Alzaabi, Fatma Khalid Badri, Kareem Darwish, Ehab Mansour Diab, Muhammad Morsy Elmallah, Amira Ayman Elnashar, Ashraf Hatim Elneima, MHD Tameem Kabbani, Nour Rabih, Ahmad Saad, Ammar Mamoun Sousou. 2022. Gulf Arabic Diacritization: Guidelines, Initial Dataset, and Results. Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), pp. 356-360.

- Ahmed Abdelali, Nadir Durrani, Cenk Demiroglu, Fahim Dalvi, Hamdy Mubarak, Kareem Darwish. 2022. NatiQ: An End-to-end Text-to-Speech System for Arabic. Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP).
- Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, Kareem Darwish. 2021 QADI: Arabic dialect identification in the wild. Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 1-10.
- Darwish, Kareem, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. 2018. Multi-Dialect Arabic POS Tagging: A CRF Approach. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). Miyazaki, Japan, May 2018.
- Darwish, Kareem, Ahmed Abdelali, Hamdy Mubarak, Younes Samih, and Mohammed Attia. 2018. Diacritization of Moroccan and Tunisian Arabic Dialects: A CRF Approach." In Proceedings of the 3rd Workshop on Open-Source Arabic Corpora and Processing Tools. Miyazaki, Japan, May 2018.
- Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki. 2017. Arabic POS Tagging: Don't Abandon Feature Engineering Just Yet. Proceedings of the Third Arabic Natural Language Processing Workshop, pp. 130--137, 2017.
- Younes Samih, Mohammed Attia, Mohamed Eldesouki, Ahmed Abdelali, Hamdy Mubarak, Laura Kallmeyer, Kareem Darwish. 2017. A Neural Architecture for Dialectal Arabic Segmentation. Proceedings of the Third Arabic Natural Language Processing Workshop, pp. 46--54, 2017.
- Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali. 2017. Arabic Diacritization: Stats, Rules, and Hacks. Proceedings of the Third Arabic Natural Language Processing Workshop, pp. 9--17, 2017.
- Hamdy Mubarak, Kareem Darwish, Walid Magdy. 2017. Abusive language detection on Arabic social media. Proceedings of the First Workshop on Abusive Language Online, pp. 52-56, 2017.
- Hamdy Mubarak and Kareem Darwish. Demographic Surveys of Arab Annotators on CrowdFlower. Weaving Relations of Trust in Crowd Work: Transparency and Reputation across Platforms Workshop, WebSci16. 2016.
- Eldesouki, Mohamed, Fahim Dalvi, Hassan Sajjad, and Kareem Darwish. "QCRI@ DSL 2016: Spoken Arabic Dialect Identification Using Textual." *VarDial 3* (2016): 221.
- Hamdy Mubarak, Kareem Darwish. Classifying Arab Names Geographically. In Proceedings of the Arabic Natural Language Processing Workshop (ANLP 2015), Beijing, China, July 2015.
- Hamdy Mubarak, Kareem Darwish, Ahmed Abdelali. QCRI@QALB-2015 Shared Task: Correction of Arabic Text for Native and Non-Native Speakers' Errors. In Proceedings of the Arabic Natural Language Processing Workshop (ANLP 2015), Beijing, China, July 2015.
- Bensalem, Imene, Imene Boukhalifa, Paolo Rosso, Lahsen Abouenour, Kareem Darwish, and Salim Chikhi. "Overview of the AraPlagDet PAN@ FIRE2015 Shared Task on Arabic Plagiarism Detection." In FIRE Workshops, pp. 111-122. 2015.
- Massimo Nicosia, Simone Filice, Alberto Barrón-Cedeño, Iman Saleh, Hamdy Mubarak, Wei Gao, Preslav Nakov, Giovanni Da San Martino, Alessandro Moschitti, Kareem Darwish, Lluís Màrquez, Shafiq Joty and Walid Magdy. QCRI: Answer Selection for Community Question Answering - Experiments for Arabic and English. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics. Denver, Colorado, June 2015.
- Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2014. Structural And Semantic Evolution Of Egyptian Political Polarization On Twitter. Qatar Foundation Annual Research Conference, SSPP0757
- Darwish K. Arabizi Detection and Conversion to Arabic. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP) 2014 Oct (pp. 217-224).
- Hamdy Mubarak , Kareem Darwish. Using Twitter to Collect a Multi-Dialectal Corpus of Arabic. In Proceedings of the Arabic Natural Language Processing Workshop (ANLP 2014), Doha, Qatar, October 2014.
- Hamdy Mubarak , Kareem Darwish. Automatic Correction of Arabic Text-a Cascaded Approach. In Proceedings of the workshop on Arabic Natural Language Processing Workshop (ANLP 2014), Doha, Qatar, October 2014.
- Kareem Darwish, Ahmed A. Ali, Ahmed Abdelali, 2014, Query Term Expansion by Automatic Learning of Morphological Equivalence Patterns from Wikipedia. In Proceedings of the Workshop on Semantic Matching in Information Retrieval (SMIR 2014), Gold Coast, Australia, July 2014.
- Ahmed Mourad, Kareem Darwish. 2013. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, ACL 2013, pp. 55-64.
- Saleh, I., K. Darwish, A. Fahmy. "Recognizing Multilingual Wikipedia NE's using SVM's with Threshold Adjustment", ACL 2010 NEWS workshop, 2010.

- Abdul-Hamid, A., K. Darwish. "Simplified Feature Set for Arabic Named Entity Recognition", ACL 2010 NEWS workshop, 2010.
- Darwish, K. "Transliteration Mining with Phonetic Conflation and Iterative Training", ACL 2010 NEWS workshop, 2010.
- Darwish, K. CMIC at INEX 2008: Link-the-wiki track. Advances in Focused Retrieval, 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX), Revised Selected Papers, Springer, S. Geva, J. Kamps, and A. Trotman, 2009.
- Magdy, W. and K. Darwish. Book search: indexing the valuable parts. Proceeding of the 2008 ACM workshop on Research advances in large digital book repositories, CIKM, October 30, 2008.
- Magdy, W., K. Darwish. CMIC at INEX 2007: Book search track. Focused access to XML documents, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX), Revised Selected Papers, Springer, N. Fuhr, J. Kamps, M. Lalmas, and A. Trotman, editors, Lecture Notes in Computer Science, 2008.
- Darwish, K. Searching OCR'ed Arabic documents. Workshop on Signal Processing and its Applications (WoSPA'08), 2008.
- Magdy, W., K. Darwish, O. Emam, and H. Hassan. Arabic cross-document person name normalization. Workshop on Computational Approaches to Semitic Languages, ACL, 25-32, 2007.
- Madkour, A., K. Darwish, H. Hassan, A. Hassan, and O. Emam. Bionoculars: Extracting protein-protein interactions from biomedical text. In Biological, translational, and clinical language processing: Association for Computational Linguistics, 89-96, 2007.
- Darwish, K., H. Hassan, and O. Emam. Examining the Effect of Improved Context Sensitive Morphology on Arabic Information Retrieval. In ACL Workshop on Computational Approaches to Semitic Languages, Ann Arbor, 2005.
- Darwish, K. Building a Shallow Morphological Analyzer in One Day. In ACL Workshop on Computational Approaches to Semitic Languages: 47-54, 2002.

Published, Unrefereed Papers

- Kareem Darwish. To Kavanaugh or Not to Kavanaugh: That is the Polarizing Question. arXiv preprint arXiv:1810.06687, 2018.
- Ahmed Abdelali, Mohamed Attia, Younis Samih, Kareem Darwish, Hamdy Mubarak. Diacritization of Maghrebi Arabic Sub-Dialects. arXiv preprint arXiv:1810.06619, 2018
- Mujahid Kutlu, Kareem Darwish, Tamer Elsayed. Devam vs. Tamam: 2018 Turkish Elections. arXiv preprint arXiv:1807.06655, 2018.
- Mohamed Eldesouki, Younes Samih, Ahmed Abdelali, Mohamed Attia, Hamdy Mubarak, Kareem Darwish. 2017. Arabic Multi-Dialect Segmentation: bi-LSTM-CRF vs. SVM. arXiv preprint arXiv:1708.05891, 2017.
- Magdy, Walid, and Kareem Darwish. "Trump vs. Hillary Analyzing Viral Tweets during US Presidential Elections 2016." arXiv preprint arXiv:1610.01655 (2016).
- Darwish, Kareem, and Walid Magdy. Attitudes towards Refugees in Light of the Paris Attacks. arXiv preprint arXiv:1512.04310. 2015.
- Magdy, Walid, Kareem Darwish, and Norah Abokhodair. Quantifying Public Response towards Islam on Twitter after Paris Attacks. arXiv preprint arXiv:1512.04570. 2015.
- Darwish, K. and O. Emam. Retrieving Arabic Printed Document: a Survey. ICUDL 2006.
- Darwish, K. and A. Madkour. To Part or to Whole: The GUC Goes to TREC – Genomics Track. TREC 2004.
- Darwish, K. and D. Oard. Improving OCR-Degraded Text Retrieval Using Query Translation. In Symposium on Document Image Understanding Technology: 181-187, 2003.
- Darwish, K. and D. Oard. Probabilistic Structured Query Methods. University of Maryland, College Park, UMIACS CS-TR-4457, 2003.
- Darwish, K. and D. Oard. CLIR Experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval. University of Maryland, College Park, UMIACS CS-TR-4456, 2003.
- Darwish, K. and D. Oard. CLIR Experiments at Maryland for TREC 2002: Evidence Combination for Arabic-English Retrieval. in TREC. 2002.
- Darwish, K. Building a Shallow Arabic Morphological Analyzer in One Day. University of Maryland, College Park, UMIACS CS-TR-4326, 2002.
- Darwish, K., D. Doermann, R. Jones, D. Oard, and M. Rautiainen: TREC-10 Experiments at University of Maryland CLIR and Video. TREC: 549-561, 2001.

ThinkWeek Papers

Kareem Darwish, Ahmed El-Shimi, Ahmed Abdul Hamid, Heba Ashour, Motaz El-Saban. Arabic Language Searching on the Web. 2007
Kareem Darwish, Ahmed El-Shimi, Ahmed Morsy, Hany Grees, Walid Magdy. BookWeb: Making Digitized Books browse-able. 2008
Kareem Darwish, Heba Ashour, Mona Soliman Habib. Leveraging Wikipedia for Bing. 2009
Kareem Darwish, Heba Ashour, Motaz El-Saban, Ahmed Abdul Hamid, Mona Soliman Habib. Exploiting English Resources to Enhance Bing's International Search Experience. 2009
Kareem Darwish, Ahmed Hefny, Mona Soliman Habib. Multilingual Web Search. 2010

Patents

Mubarak, Hamdy, Kareem Mohamed Darwish, Ahmed Abdelali, Hassan Sajjad, Younes Samih. Method and system for diacritizing Arabic text. Publication date: 2022/6/16, US Patent Office, Application number: 17598633.
Magdy, Walid, Kareem Darwish, Ahmed Ali. System and method for automatic generation of information-rich content from multiple microblogs, each microblog containing only sparse information. US Patent Office, Publication date: 2018/6/5
Patent number: 9990368, Application number: 14408732
Ali, Ahmed and K. Darwish. Automated Admission. US Patent Application. Filed February 9, 2012. Application serial No. 13/370,097
Darwish, K. Online Communities. Filed August 26, 2011. Application serial No. 13/218,828
Abdul-Hamid, A. and K. Darwish. Translation of multilingual embedded phrases. US Patent Application. Publication date: 2015/8/4, US Patent Office, Patent number: 9098488, Application number: 13078987
Darwish, K. Choosing the next document. US Patent Application. Filed Dec. 23, 2008.
Darwish, K. and A. Morsy. Method for making digital documents browseable. US Patent Application 2009/0083677 A1.
Morsy, A. and K. Darwish. Summarizing documents with marked points. US Patent Application 2009/0083026 A1.
Darwish, K. and O. Emam. Method and system for providing information to transportation vehicle on the presence of passengers. US Patent Application 2009/0128362.

Skills

Programming Languages & Systems: Python, Perl, C#, C++, Java, Linux, Windows

Languages

Fluent in Arabic and English and a speaker of Hebrew and Turkish

Citizenship

USA

Awards Received

- Distinguished Creator of the Year, 2020, Qatar Foundation
- QCRI Thanaa Award 2012.
- At Microsoft:
 - 2 Microsoft gold star awards
 - 3 years 20% contribution ranking
 - 2 years exceeded commitment rating
- Deans List at UMCP and UMBC
- Most Outstanding Freshman Engineering Student Award (UMBC)
- Presidential Merit Scholarship (UMBC)
- Diane Berman Merit Award (UMCP)

References Available upon Request