Research Statement

Mossaab Bagdouri, Ph.D. candidate in Computer Science, University of Maryland, mossaab@umd.edu

My primary research interest revolves around developing better information retrieval systems. I address this goal from three broad perspectives. First, I study techniques for enhancing the effectiveness of Information Retrieval (IR) systems leveraging machine learning algorithms. Second, I investigate methods for improving the efficacy of IR evaluation using statistical methods. Third, I make use of the abundant availability of "Big Data" to increase the range of information needs that can be addressed and to generate several types of useful insights. I find my passion in introducing new research problems. I then address these novel problems with classic solutions where possible, or I propose novel solutions (see Table 1).

Enhancing the Effectiveness of Information Retrieval Systems

Today's search engines rarely limit themselves to hand-built scoring functions to rank the documents to be presented to a user. Typically, a variety of machine learning techniques are applied at various stages to improve the results. For instance, unsupervised algorithms can be used to partition the search space or to remove duplicate documents from result sets before presentation to the user. Supervised algorithms, for example, can be used to suppress spam or to re-rank the final set of documents. I study a diverse set of these algorithms in multiple applications.

I have studied in my dissertation how deep neural networks can accurately identify questions asked on microblogs that convey real information needs. Question marks, for instance, might incorrectly detect rhetorical questions, so solving this problem is an important stage for building agents that proactively answer questions occurring naturally in a conversation. I found that Bidirectional Long Short-Term Memory (BLSTM) neural networks substantially improve the performance over two previous state-of-the-art techniques, both based on Support Vector Machines. A similar neural architecture can also be applied to detect answerable questions (i.e., those for which sufficient context is included within the question so that some stranger, such as an expert, might be able to provide a useful answer) [2].

Fewer than one in four answerable questions asked on Twitter receive useful answers. This is an enormous opportunity for increasing user satisfaction. I have studied how old user-generated answers, such as those posted on Yahoo! Answers, can be leveraged to answer Twitter questions. I examined some transformations that accommodate differences in writing conventions between those two services, and found that when used with learning-to-rank techniques retrieval effectiveness from Yahoo! Answers can be substantially increased [1].

Applications of classification algorithms for improving document retrieval are endless. For example, I have used them to study how to find professionals in Twitter [7], how to predict the deletion of tweets [6], and how to retrieve informal content in Arabic forums for an English query [8]. In the future, I'm interested in examining algorithms for ranking documents from sources with distinctive characteristics (e.g., microblogs, encyclopedias and news articles). I want to study how we can "translate" these heterogeneous documents into some unified representation, to allow cross-platform document scoring.

Table 1: The space of IR topics I study is dominated by new research problems, which I address with both classic and novel solutions.

	Old Techniques	New Techniques
New Problems	• Detecting answerable questions in microblogs [3] • Answering Twitter questions with Yahoo! Answers [2] • Profession-based person search in microblogs [7] • Predicting deletions of microblog posts [6] • Cross-language information retrieval for informal content [8]	• Minimizing the annotation cost of certified text classification [9, 11] • Evaluation of systems retrieving only new documents [4]
Old Problems	• Journalists and Twitter: Description of usage patterns [1] • News aggregator in the MENA region [12]	• Pearson rank [10] • Live question answering [5]

Research Statement

Mossaab Bagdouri, PhD candidate in Computer Science, University of Maryland, mossaab@umd.edu

Improving the Efficiency of Information Retrieval Evaluation

Retrieval and classification systems can be improved only if we can reliably measure their performance. This evaluation might be affected by several factors, such as constraints on the annotation budget, and non-reusability of available test collections. I addressed these challenges in two lines of work.

For developing a text classifier with a suitable target on its effectiveness, it is a widespread practice to add training documents sequentially, until that goal is met. This process, however, introduces a sequential test bias that is unacceptable when the effectiveness of the classifier must be certified (as is often the case in the domain of legal search) [11]. This is because we are more likely to stop training when we overestimate the effectiveness of the classifier than when we underestimate it. This problem becomes more challenging with a limited annotation budget, when we also need to balance between the documents allocated for training (to improve the performance of the classifier) and those dedicated for testing (to increase our confidence on measuring that performance). Based on simulation-based power analysis methods, I introduced a framework for minimizing these two sets without sacrificing the validity of this certification [9]. For a predetermined certifiable effectiveness, this framework estimates the size of the two sets during training (through n-fold cross-validation within the growing training set), before testing the classifier, only once, on a held-out test set of a learned size. While policies that use this framework have yet to be developed, experimental evidence shows a potential for saving as much as 40% of the annotation budget.

The cost for gathering annotations to measure the performance of new retrieval algorithms is typically amortized by releasing these annotations as test collections for future use by other researchers. But this is beneficial only when those test collections are reusable. In traditional information retrieval evaluation campaigns, a collection of documents and a set of topics are distributed to the participants of a shared task who return a list of potentially relevant documents. The assessments indicate the relevance of a subset of <document, topic> pairs, and future systems (i.e., those that did not participate in that campaign), can use the labels of these pairs to evaluate their performance. In the recent Live Question Answering (LiveQA) evaluation campaign organized by the National Institute of Standards and Technology, no document is distributed, and systems can use whatever resources to return answers to real-time questions. This created a challenge for reusing the annotations by future systems. In fact, if a future system returns answers only from a new resource (e.g., future articles), then there is no obvious way for assessing the effectiveness of that system on those questions. This problem motivated my work on estimating the relevance of new answers based on their similarity with previously labeled answers [4]. The suggested approach, based on the representation of words as dense vectors, computes a vector of the "core" elements of an answer by subtracting the average vector of "bad" answers from the average vector of "good" answers. The cosine similarity of the new answer (also in a dense representation) determines the estimate of its relevance. I showed through two ablation studies that this approach preserves the distance between top systems [10], and succeeds even in detecting future best systems.

Can we minimize the annotation budget by both balancing between training and test and estimating the relevance of unassessed documents? This problem has been studied in the subfield of machine learning known as active learning, but with a focus on training documents. In future work, I want to examine the effect of active selection of test documents on the ability to reliably evaluate classification systems.

Research Statement

Mossaab Bagdouri, PhD candidate in Computer Science, University of Maryland, mossaab@umd.edu

Better Retrieval and Useful Insights with Big Data

"Free" Big Data can sometimes contribute to improving the effectiveness of retrieval systems. The work discussed earlier on answering Twitter questions with Yahoo! Answers has been made possible after I collected a large crawl of 260 million questions and 1.4 billion answers from Yahoo! Answers. That same collection demonstrated its usefulness in the context of LiveQA in two ways [5]. First, the size of the crawl increased the chances of finding old questions similar to new questions, and for which the answers could be useful as well. Second, I could exploit user-indicated ratings of answers to train a deep neural network for learning to rank answers. As a consequence, my system scored the highest among all automatic systems that participated in LiveQA.

Another usage of Big Data is for obtaining insights about a particular population of users, which might help designing better tools for them. Last year, for instance, I mined over 2 billion tweets on a Hadoop cluster to analyze how news producers and consumers use Twitter [1]. In that work, I discovered several insights, such as that the user behavior of journalists depends on their media type (i.e., print, radio or television), and that Arabic speaking journalists use Twitter for disseminating news, while journalists in English speaking countries engage their audience by having a two-way communication. With those findings, one might suggest that the tools that would support Arab journalists in their dissemination role should be different from those of the English journalists as they seek information using Twitter.

Observing that major news aggregators have a low coverage in my home country (Morocco), I developed an automated news aggregator that gathers stories from over a hundred of online Moroccan newspapers, classifies their content before ranking them and displaying the top stories on the main page. The success of this project led to the launch of a series of similar websites for six other Arab countries that attract, combined, more than 100,000 unique visitors every day. The 40+ million Arabic news articles I have collected for a period spanning over a decade can help understand the major events that the MENA region has gone through recently, especially with respect to the "Arab spring."

References

[1] **Bagdouri, Mossaab**. Journalists and Twitter: A multidimensional quantitative description of usage patterns. In: ICWSM. Cologne, Germany, 2016, pp.22–31.

[2] **Bagdouri, Mossaab** and Douglas W. Oard. Building Bridges across Social Platforms: Answering Twitter Questions with Yahoo! Answers. In: SIGIR. Tokyo, Japan, 2017. To appear.

[3] **Bagdouri, Mossaab** and Douglas W. Oard. Detecting Answerable Questions in Microblogs. In preparation.

[4] **Bagdouri, Mossaab** and Douglas W. Oard. On the Reusability of Open-Resource Test Collections: Estimating Relevance with Word Embeddings. In preparation.

[5] **Bagdouri, Mossaab** and Douglas W. Oard. CLIP at TREC 2016: LiveQA and RTS. In: TREC. Gaithersburg, MD, USA, 2016.

[6] **Bagdouri, Mossaab** and Douglas W. Oard. On predicting deletions of microblog posts. In: CIKM, Melbourne, Australia, 2015, pp.1707–1710.

[7] **Bagdouri, Mossaab** and Douglas W. Oard. Profession-based person search in microblogs: Using seed sets to find journalists. In: CIKM. Melbourne, Australia, pp.593–602.

[8] **Bagdouri, Mossaab**, Douglas. W. Oard, and Vittorio Castelli. CLIR for informal content in Arabic forum posts. In: CIKM. Shanghai, China, 2014, pp.1811–1814.

[9] **Bagdouri, Mossaab**, William Webber, David D. Lewis, and D. W. Oard. Towards Minimizing the Annotation Cost of Certified Text Classification. In: CIKM. San Fransisco, CA, USA, 2013, pp.933–936.

[10] Gao, Ning, **Mossaab Bagdouri**, and Douglas W. Oard. Pearson Rank: A Head-Weighted Gap-Sensitive Score-Based Correlation Coefficient. In: SIGIR. Pisa, Italy, 2016, pp.941–944.

[11] Webber, W., **Mossaab Bagdouri**, David D. Lewis, and Douglas W. Oard. Sequential Testing in Classifier Evaluation Yields Biased Estimates of Effectiveness. In: SIGIR. Dublin, Ireland, 2013, pp.989–99

[12] www.maghress.com