MP 06W0000108

MITRE PRODUCT

A Methodology for End-to-End Evaluation of Arabic Document Image Processing Software

June 2006

Paul M. Herceg
Catherine N. Ball

MITRE
Washington C3 Center
McLean, Virginia

Abstract

This paper describes a methodology for end-to-end evaluation of Arabic document image processing software. The methodology can be easily tailored to other languages, and to other document formats (e.g., audio and video). Real-world documents often involve complexities such as multiple languages, handwriting, logos, signatures, pictures, and noise introduced by document aging, reproduction, or exposure to environment factors. Information retrieval systems that implement algorithms to account for such factors are maturing. The proposed methodology is vital for measuring system performance and comparing relative merits.

Introduction

Various software solutions have been proposed for digitization and understanding of noisy, complex Arabic document images. Optical-character-recognition-based (OCR-based) solutions have been available for decades; however this technology is often tailored to the most common document image type: clean, monolingual documents. Real-world documents often involve multiple languages, handwriting, logos, signatures, pictures, stylized text, and other document aspects. Real-world documents involve noise introduced by document aging, reproduction, or exposure to environment factors. Document image processing solutions are maturing to deal with such complexities. Such systems include image clean-up algorithms and page segmentation, followed by various recognition or digitization algorithms: OCR, handwritten word recognition (HWR), logo identification, signature identification, sub-image or picture identification. Indexing digitized document renditions into a search engine enables ad hoc querying of the collection. Some researchers have proposed semi-automation, a process in which human readers interpret complex documents and record a spoken rendition; the audio recordings are then processed by a spoken document retrieval (SDR) system [Garofolo00], employing automatic speech recognition (ASR) for digitization and an information retrieval solution to enable ad hoc queries. To handle foreign language, machine translation may be included in any of the aforementioned document image processing systems. This array of approaches results in widely varying performance. This paper discusses a methodology for evaluating the end-to-end retrieval performance of these systems.

End-to-end measures, rather than component measures, are necessary for quantifying end-to-end retrieval performance (i.e., search); it is insufficient to rely on component measures to assess the merit of an integrated system for document image processing. In other words, the integration of best-of-breed components does not necessarily result in the best end-to-end performance. For example, one might identify the best performing page segmentation algorithm, OCR engine, and search engine; however, the resulting end-to-end document image processing system may measure surprisingly below a comparable end-to-end system.

In addition to performance variations, the techniques and system architectures for dealing with document image complexity can vary widely. An end-to-end evaluation methodology enables the inputs and outputs to be homogenized for a fair comparison of very different system implementations—for example, comparing OCR against HWR, and those against systems that use a human-in-the-loop for preprocessing.

The Ad Hoc Use Case

A variety of use cases[1] has been applied for evaluating document image processing software; one of these—the *ad hoc use case*—is the focus of the methodology discussed here. The first eight Text REtrieval Conferences (TREC) hosted by the National Institute of Standards and Technology (NIST) defined the ad hoc task as an information retrieval use case where the document collection is fixed, users submit queries to the information retrieval system, and the system returns a set of ranked retrieval results (i.e., documents) [Harman05, p.80]. In the ad hoc use case, the rank of a retrieved document is associated with a score of relevancy to the query. The ranked retrieval results are useful if the user needs information on a given topic, in which case only 10−100 results may be reviewed. The results can also be used to separate a document collection into relevant and nonrelevant documents for a given query or set of queries. In this case, if the collection is large, and a significant portion is relevant to a submitted query, ranked retrieval results can become quite large.

The query in the ad hoc use case could be represented in a variety of forms. The conventional query format is electronic text. However, for document images, the user may be interested in searching for a sample image, which could be a signature, logo, photograph, or another image. Some emerging commercial systems use sample-image queries.

Measures

More than a decade of TREC literature provides guidance on measuring the ad hoc use case. Measures have been debated and have evolved over time, and widely accepted overall system performance measures have emerged. Because "comparisons among TREC systems are most often made in terms of [mean average precision] MAP, R-precision, or precision at a small document cutoff level such as P(10) or P(30)" [Buckley05, p.67], the methodology discussed here focuses on the overall system performance reflected in MAP and average R-precision measures, both of which are defined below.

These overall system measures are based on *precision* and *recall*, which measure the effectiveness of retrieval results associated with a given query. Precision and recall depend on the concept of document relevancy. The specific document relevancy approach used in this methodology is discussed in the following section.

Precision and recall for a given query and associated retrieval results are widely documented in information retrieval literature. Recall is the proportion of true positives to the expected retrieval results (i.e., to the sum of the true positives and true negatives). Precision is the proportion of true positives to the retrieval results (i.e., to the sum of the true positives and false positives).

Mean average precision [Harman94] is a collective information retrieval measure that reflects both precision and recall. This is important because the ideal system has both high precision and high recall. This collective measure is based on *non-interpolated average precision* calculations. Manning and Schütze describe non-interpolated average precision of retrieval results as follows:

[1] "Actors represent the people or things that interact in some way with the system...Use cases represent the things of value that the system performs for its actors." [Bittner03, p.3]

> Uninterpolated [or non-interpolated] average precision aggregates many precision numbers into one evaluation figure. Precision is computed for each point in the list where we find a relevant document and these precision numbers are then averaged…. Precision at relevant documents that are not in the returned set is assumed to be zero. This shows that average precision indirectly measures *recall*, the percentage of relevant documents that were returned in the retrieved set (since omitted documents are entered as zero precision)…. Average precision is one way of computing a measure that captures both precision and recall. [Manning99, pp. 535-6]

For a query set, each query produces a ranked retrieval result set with an associated non-interpolated average precision value. The mean average precision is the mean over all non-interpolated values, keeping in mind that queries for which there are no retrieval results contribute zero to the mean.

Average R-precision is yet another collective information retrieval measure that reflects both precision and recall. It is derived from individual R-precision calculations. The R-precision of a query's retrieval results is the precision at rank R, where R is the cardinality of the expected retrieval results. For a query set, each query produces a ranked retrieval result set with an associated R-precision value. The average R-precision over all queries is the mean over all values, again keeping in mind that queries for which there are no retrieval results contribute zero to the mean.

Methodology

Essentially the document image processing software evaluation methodology involves submitting a set of queries to a system and analyzing the retrieval results and TREC measures. This methodology applies to systems that digitize document images and produce retrieval results according to the ad hoc use case. Document image processing systems that solely convert documents to a digital form (e.g., electronic text) are not covered by this methodology; however, such systems are easily extended by adding an information retrieval component. The implementation of this methodology follows and involves open source and custom-developed tools.

The first step toward applying the TREC measures is to define the concept of document relevancy. For this methodology, a document is relevant to a query if the query expression exists in the document image ground truth.

The TREC evaluations were based on "pooling" retrieval results from the multiple systems participating at a TREC event in order to establish the set of ground truth relevant documents for each query. This approach was fine for comparing the relative merits of a closed set of systems at a given point in time (i.e., those participating at a given TREC event). However, this approach is not appropriate for comparing systems that are identified over a period of time, particularly because the full set of true negatives for the retrieval results set is unknown.

Contrary to the TREC approach, in this methodology, human-developed ground truth for each document image enables TREC measures to reflect the entire set of true negatives. As a result, any number of systems can be evaluated against a given ground truth set. Systems to be evaluated can be identified over time, and the performance of each can be compared with prior evaluation runs.

For this methodology, the query set of interest is the entity types handled by information extraction tools reviewed at the Message Understanding Conferences (MUC), also hosted by NIST since 1990. Therefore, MUC-7 entities [Chinchor97] are the query expressions. The list of MUC-7 entities occurring in a document image is called the *extraction ground truth*. Extraction ground truth is derived from keyboarded transcriptions and translations.

For the purpose of discussion, the document image processing software to be tested is called the *tool under test*. The method for evaluating the tool under test includes the following steps and is depicted in Figure 1.

1. Ingest the document images to the tool under test.
2. Transform the extraction ground truth for each document image into a set of queries and a set of expected retrieval results for each query.
3. Submit the queries to the tool under test.
4. Aggregate the retrieval results over all queries.
5. Calculate the overall system measures by comparing the retrieval results to the expected retrieval results.
6. Transform the trec_eval output and correlate scores with the ground truth.
7. Analyze the data.

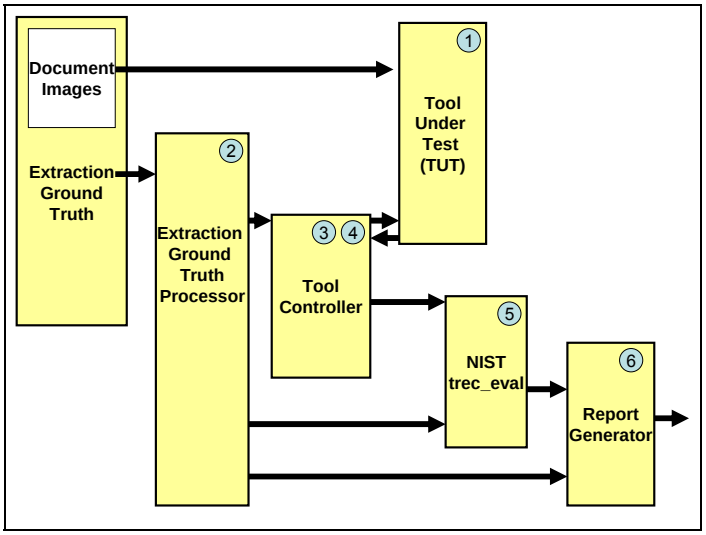

Figure 1. Producing Evaluation Data.

Buckley and Voorhees state that the NIST "*trec_eval* program… is publicly available and has become the primary method for evaluating retrieval results" [Buckley05, p.53]. See [trecevalnd]. This established tool is re-used here and obviates the need to develop a new measurement tool. The *extraction ground truth processor* and the *tool controller* ensure that the expected retrieval results and the aggregated retrieval results match the file formats required by trec_eval.

The data analysis includes comparing MAP and average R-precision measurements, and identifying the features of problematic queries.[2] This methodology was initially applied to a very

[2] Future analysis will include additional feature data about each query: script type, glyph type (e.g., machine text, handwritten text, logo text, signature text, stamp text), page attributes.

small data set. Larger data sets are planned, for which more extensive data analysis will be performed.

Applicability to Other Types of Collections

The prior section discussed a generalized methodology for measuring the ad hoc use case for document image processing systems in order to compare and contrast system performance.

This methodology can be applied to a variety of collection types. The extraction ground truth can be tailored to query other document content types, for example pictures and logos. Also, this approach could be used to measure the performance of spoken document retrieval systems or video retrieval systems. In these cases, the user begins with a collection of audio files or a collection of video files respectively, rather than document images. The audio files or video files are ingested by the tool under test. Queries could be electronic text, audio clips (for audio or video), or sub-images (for video), depending on the capabilities of the information retrieval software.

Upon completion of scheduled evaluations with this methodology, the authors plan to tailor the system and extraction ground truth to audio and video evaluation.

References

[Bittner03] Bittner, K., Spence, I. *Use Case Modeling*. Addison-Wesley, 2003.

[Buckley05] Buckley, C. and Voorhees, E. M., "Retrieval System Evaluation." In Voorhees, E. M. and Harman, D. K. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, Cambridge, MA. 2005, pp. 53-75.

[Chinchor97] Chinchor, N. "MUC-7 Named Entity Task Definition, Version 3.5". http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html. 17 September 1997.

[Garofolo00] Garofolo, J. S., Auzanne, C.G. P., Voorhees, E. M. "The TREC Spoken Document Retrieval Track: A Success Story." In Proceedings of RIAO 2000, April 2000.

[Harman94] Harman, D. K. "Overview of the Second Text REtrieval Conference." In The Second Text REtrieval Conference, pages 1-20. National Institute of Standards and Technology. 1994.

[Harman05] Harman, D. K. "The TREC Ad Hoc Experiments." In Voorhees, E. M. and Harman, D. K. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, Cambridge, MA. 2005. pp. 79-97.

[Manning99] Manning, C. D., Schütze H. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA. 1999.

[trecevalnd] trec_eval 7.3, http://trec.nist.gov/results.html