Evaluating the Importance of Image-related Text for Ad-hoc and Case-based Biomedical Article Retrieval

Matthew S. Simpson, MSc; Dina Demner-Fushman, MD, PhD; and George R. Thoma, PhD

Lister Hill National Center for Biomedical Communications (LHNCBC)
U.S. National Library of Medicine, Bethesda, MD 20894

ABSTRACT

Images and their associated text are an essential source of information in biomedical articles. However, their use in providing evidence for clinical case descriptions has yet to be evaluated in the context of information retrieval. Given the complexity of case-based document retrieval, understanding the importance of images and image-related text is critical for future research into text- and content-based approaches to this problem. In this study, we compare the extent to which image-related text is useful in facilitating document retrieval for both case-based information requests and ad-hoc clinical questions in the domain of family practice. We show that case-based document retrieval is significantly improved with the use of image-related text whereas retrieval for clinical questions is largely unaffected. This suggests that visual evidence is more relevant for the case descriptions used in our study than the clinical questions.

INTRODUCTION

Images contained in peer-reviewed biomedical articles are an important form of *visual evidence* and provide essential information for researchers and clinicians. Figures and tables, along with their associated captions, are invaluable for use in education, clinical decision support, and evidence-based approaches to patient care. However, biomedical images are seldom self-evident, and much of the information required for their comprehension is found elsewhere in an article: captions, titles, abstracts and snippets of text within the body of an article that mention an image (mentions) are all integral to image understanding.[1] As such, the retrieval of images from biomedical articles based on their visual properties and associated text remains an actively researched topic.[2]

In this study, we report on the importance of image-related text (i.e., the captions and mentions of both figures and tables) in two information retrieval scenarios in the domain of family practice. In the first, we evaluate the usefulness of image-related text in providing evidence related to *ad-hoc* clinical questions. These information requests are typically brief and address a specific need, such as "what is the best way to evaluate an acute traumatic knee injury?"[3] In the second scenario, we seek to determine the extent to which image-related text can provide evidence for *case-based* information needs. A clinical case description is typically a detailed report of the symptoms, treatment, and demographic profile of an individual patient and often describes an unusual or novel occurrence.

In our experiment, we utilized techniques of natural language processing (NLP) to automatically construct a large test corpus of biomedical articles and generate a set of batch queries for both the ad-hoc and case-based retrieval scenarios. Mapping article text to concepts in the Unified Medical Language System® (UMLS®) Metathesaurus®[4] was instrumental in the success of our query generation and search strategies.

Our most significant finding was that including image captions and mentions in our document representation significantly improved retrieval for the case-based information requests but not for the ad-hoc clinical questions. Therefore, we conclude that visual evidence is more relevant to the case descriptions used in our evaluation than the ad-hoc clinical questions.

BACKGROUND

ImageCLEF, part of the Cross Language Evaluation Forum (CLEF), is a working conference for researchers to discuss and evaluate cross-language image retrieval methods. For the medical retrieval task, participants are given a set of multilingual ad-hoc information requests (topics) and are tasked with retrieving the most relevant images for each topic from a large collection of biomedical articles.

Text-based approaches to this problem have been well-studied in the past, and a number of retrieval sys-

tems have been developed for retrieving images from biomedical articles—including the Yale Image Finder (YIF)[5] and the GoldMiner[6] and BioText[7] search engines. In addition to image captions, current systems commonly rely on an article's title and abstract text to retrieve the images contained therein.

At ImageCLEFmed'09,[2] case-based topics were introduced. In this task, participants were given clinical case descriptions and asked to retrieve the most relevant related cases and supporting articles. In our approach[8] to the case-based task, we represented the articles in the collection by their title, abstract, Medical Subject Headings (MeSH® terms), and the captions and mentions from all their contained images. We then used UMLS concepts extracted from the case descriptions to form queries and retrieve the relevant documents. Our system achieved a Mean Average Precision (MAP) of 0.34 for the case-based topics, which was the highest at the meeting.

In our current study, our goal is to better understand the utility of incorporating image-related text into our document representation. While our particular emphasis is on improving case-based document retrieval, we are also interested in determining whether image-related text benefits ad-hoc document retrieval (as apposed to image retrieval) tasks as well.

METHODS

We focused our experiment on articles from the *Journal of Family Practice* "Photo Rounds" (PR) and "Clinical Inquiries" (CI) features. A PR article typically presents a detailed description of a clinical case, along with relevant images, and then describes a differential diagnosis while providing supporting evidence with references. Similarly, A CI article usually poses a brief clinical question and then summarizes an evidence-based answer using knowledge from supporting references. For example, while a PR article might describe a patient with symptoms similar to pigmented Bowen's disease[9], a CI article might ask for the most effective treatment of the disease.

With these two collections, we can compare the importance of image-related text in providing evidence for the case descriptions contained in the PR articles versus the ad-hoc clinical questions in the CI articles. Crucially, for each article we assume its cited references to be the set of relevant documents. The remainder of this section describes (1) the creation of our test corpus, (2) our document representations, (3) the creation of our topics, (4) strategies for generating queries from the topics, and (5) our evaluation.

Source	Articles
American Journal of Public Health	589
Annals of Family Medicine	129
Antimicrobial Agents and Chemotherapy	1411
Archives of Disease in Childhood	347
BMJ	331
Gut	353
Heart	441
Radiographics	1285
Radiology	4421
Thorax	308
Total	9561

Table 1: The number of articles per source in the test corpus.

Corpus Creation

To obtain the *relevant documents* for our test corpus, we began by downloading the full text HTML PR and CI articles from the website of the *Journal of Family Practice*. We then parsed the HTML documents and extracted the list of references from each article. We used the Entrez ESearch utility to find PubMed® Identifiers (PMIDs) of as many references as possible, and then downloaded the articles, using the ELink utility to obtain the primary LinkOut provider for each PMID, and added them to the corpus. Most of the full text articles were HTML documents, but many were PDF. We used the `pdftotext` program to convert the downloaded PDF documents to plain text when possible.

To introduce noise into our document collection, we added numerous other articles from various sources to the corpus with *no relevance judgements*. Table 1 summarizes the articles obtained from each additional source. With the exception of the *Radiology* and *Radiographics* journals, we downloaded from the journals' websites the full text HTML of all articles from the two most recent complete years of publication (2008–2009). The remaining articles were obtained through participation in ImageCLEFmed'09.[2]

Surrogate Documents

In order to effectively represent the articles in the corpus, we created two surrogate document formats: one with and the other without image-related text. *Citation documents* contain an article's title, abstract, and MeSH terms, and *image documents* contain the captions and mentions of all the figures and tables from an article in addition to it's citation elements.

For the citation elements, we used the Entrez EFetch utility to obtain the titles, abstracts, and MeSH terms of each article from MEDLINE®. For the image-

related elements, we first used the sentence boundary detection method described by Kiss and Strunk [10] to divide the full text articles into sentences, and we then used regular expressions to extract the captions and mentions. Sentence-level tokenization was particularly helpful for identifying captions and mentions in the converted PDF documents since, unlike the HTML documents, they lack any formatting markups. Note that while our approach to caption and mention extraction relies mostly on pattern matching, more advanced approaches [11] exist for identifying text that summarizes biomedical image content in general.

We separately indexed the citation documents and image documents with the Essie [12] search engine. Essie is particularly well-suited for information retrieval tasks in the medical domain since it performs concept-based indexing, [13] automatically expands query terms along synonymy relationships in the UMLS Metathesaurus, and weights term occurrences according to their document location when computing document scores. For example, we weight term occurrences in titles more so than those in abstracts or MeSH terms.

Structured Topics

We created structured topics for the information requests by first discarding all articles for which we were unable to obtain at least two references when creating the test corpus, and then extracting the pertinent text from an equal number of the remaining PR and CI articles. The case-descriptions from PR articles and the clinical questions from the CI articles were easily identifiable by their HTML markups.

We then followed the method described by Demner-Fushman and Lin [14] to extract PICO elements from the text of the information requests. PICO is a mnemonic for the four components of a well-formed clinical question: Patient/Problem, Intervention, Comparison, and Outcome. To construct the PICO frames, we used Essie to map the text to the UMLS Metathesaurus and then extracted concepts relating to *problems*, *interventions*, *age*, *anatomy* and *drugs*. The problem, age, and anatomy terms contribute to the Patient/Problem element of the PICO frame and the intervention and drug terms contribute to the Intervention and Comparison elements. An additional extractor identified terms related to image *modality*. Finally, we used the NegEx algorithm [15] to identify negated terms.

For example, Figure 4.3 shows the case description and PICO elements extracted from the "Photo Rounds" article entitled *Growing plaque on foot*, [9] represented as an Extensible Markup Language (XML) document. Note that the extracted terms are grouped by sentence, and within each, the terms are tagged with their corresponding PICO element types. Note also that only two sentence groups are shown for brevity.

```
<case id="6647">
 <description>
  An 82-year-old African American woman with a history of pancreatic can-
  cer came into the clinic for evaluation of a growing, asymptomatic lesion
  on her right dorsal foot. She first noticed the lesion a year ago, when it
  was pinpoint size. It was now a 2.5 cm × 1 cm hyperpigmented plaque.
  The lesion was dark brown and black and had irregular borders. It also
  had a central hyperkeratotic area (Figure 1). There was no inguinal lym-
  phadenopathy. We performed an incisional biopsy.
 </description>
 <pico>
  <sentence number="0">
   <anatomy cui="C0016504" negstatus="not_negated">
    feet
   </anatomy>
   <problem cui="C0030297" negstatus="not_negated">
    pancreatic cancer
   </problem>
   <age>
    82-year-old, 80+ years
   </age>
  </sentence>
  <!-- sentences 1 through 4 not shown -->
  <sentence number="5">
   <problem cui="C0578736" negstatus="negated">
    inguinal lymphadenopathy
   </problem>
  </sentence>
  <!-- sentence 6 not shown -->
 </pico>
</case>
```

Figure 1: Example topic generated from a *Journal of Family Practice* "Photo Rounds" article.[9]

Queries

We generated queries from the structured topics following two strategies. For the *concept-based* strategy, we composed a query as the disjunction (`OR`) of all non-negated and possibly negated terms in the topic's PICO frames. We eliminated all terms that were definitely negated. For the *type-based* strategy, we first organized the terms by type (i.e., problem, intervention, etc.), and followed the concept-based strategy to create a disjunction of terms for each type. We then formed the query as the conjunction (`AND`) of the resulting type groups. For example: `(problem_1 OR problem_2) AND (intervention_1 OR intervention_2)`.

Evaluation

In evaluating the importance of imaged-related text, we sought to determine (1) whether the inclusion of image-related text improves document retrieval for both case-based topics and ad-hoc clinical questions and (2) if the concept- and type-based queries produce significantly different retrieval results for each set of topics. Therefore, we performed 8 batch retrieval runs (2 topic sets × 2 document representations × 2 query generation strategies) over our test corpus.

Data Element		Photo Rounds	Clinical Inquiries
Topics	Total	50	50
	Extracted Terms	1091	65
References	Total	212	232
	Extracted Captions	637	661
	Extracted Mentions	1169	1363

Table 2: Summary of data elements obtained from the *Journal of Family Practice*.

Since the number of documents judged relevant is small in comparison to the size of the corpus, we used the binary preference (bpref) retrieval evaluation metric computed by `trec_eval`, which is more robust than mean average precision (MAP) when given incomplete relevance judgements.[16]

We followed the method outlined by Smucker et al. [17] to compute two-sided Fisher randomization tests in order to measure the statistical significance of our retrieval results. The randomization (or permutation) test is more reliable than the Wilcoxon signed-rank test and more general than the paired Student's t-test.

RESULTS

In this section, we describe results related to our test corpus and the batch retrieval of articles for both the case-based topics and ad-hoc clinical questions.

Corpus Creation

Table 2 summarizes the data we obtained from the *Journal of Family Practice*, which served as our source of information requests (topics) and relevant articles (references). We obtained at least two references for 50 PR articles, and from their case descriptions, extracted 1091 terms (avg. 22 terms per topic). We obtained 212 articles referenced by the 50 PR articles, averaging 4 relevant documents per topic. Finally, we identified and extracted 637 captions and 1169 mentions from the referenced articles (avg. 3 captions and 6 mentions per reference).

To equal the number of case-based topics, we included in the set of ad-hoc clinical questions the 50 most recent CI articles for which we obtained at least two references. Since these articles begin with a brief question rather than a full case description, only 65 terms could be extracted (avg. 1 term per topic). We obtained 232 references for the CI articles (avg. 5 relevant articles per topic), and from these references identified and extracted 661 captions and 1363 mentions (avg. 3 captions and 6 mentions per reference).

Retrieval

Table 3 summarizes the batch retrieval results for the 50 case-based topics (PR) and 50 ad-hoc clinical questions (CI). For each topic group, the average bpref is given for the concept- and type-based queries on both the citation documents and image documents.

For the case-based topics, the use of the concept-based query generation strategy resulted in a substantially higher average bpref than did the type-based strategy. Most notably, the average bpref on the image documents (0.7379) was a 10% increase (δ) over the average bpref on the citation documents (0.6684) at the 0.0002 significance level (p). However, for the ad-hoc clinical questions, the concept- and type-based query strategies resulted in nearly identical average bpref scores, and there was no statistically significant difference in bpref with the inclusion of image-related text.

DISCUSSION

The most significant result from our evaluation was that the use of image-related text significantly improved document retrieval for the case-based topics, but had little effect for the clinical questions. This is especially interesting since the relevant document set of the clinical questions contained more image-related text than that of the case descriptions. Therefore, visual evidence was more relevant for the case descriptions than for the ad-hoc clinical questions. This suggests that image-related text should be utilized to improve retrieval accuracy for case-based information needs, particularly in the domain of family practice.

Additionally, this result indicates that combining textual approaches with techniques from content based image retrieval (CBIR) is a promising direction for further improving retrieval for case-based topics. For example, the output of our text-based image modality extractor could be used with a content-based image modality classifier[18] to retrieve or re-rank documents according to a particular image type.

The second significant finding from our evaluation was that the concept-based query strategy achieved better retrieval results than did the type-based strategy, especially for the term-rich case-based topics. Since there was only one term extracted on average for each clinical question, there was essentially no difference between the concept- and type-based queries for these topics. This indicates the occurrence of a few specific terms is more important for document relevance than the co-occurrence of several less specific terms. Determining better query strategies is a possible avenue for future work related to case-based retrieval.

Topic Group	Queries	bpref		δ (%)	p
		Citation Documents	Image Documents		
Photo Rounds	concept-based	0.6684	0.7379	10.40	0.0002
	type-based	0.0040	0.0153	280	0.5027
Clinical Inquiries	concept-based	0.7837	0.7932	1.212	0.5004
	type-based	0.6888	0.7072	2.671	0.1232

Table 3: Retrieval results for the "Photo Rounds" and "Clinical Inquiries" topics.

CONCLUSION

Images are an important source of visual information contained in biomedical articles. In this study, we compared the extent to which image-related text is useful in facilitating document retrieval for both case-based information requests and ad-hoc clinical questions in the context of family practice. We found that considering image-related text—in addition to article titles, abstracts, and MeSH terms—significantly improved case-based document retrieval whereas retrieval for clinical questions in the same domain were largely unaffected. This suggests a promising direction for improving case-based document retrieval through the use of both text- and content-based methods.

Acknowledgements

The authors thank their colleague Dr. Charles Sneiderman for discussions of resources relevant to family practice. This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and Lister Hill National Center for Biomedical Communications.

REFERENCES

1. Yu H, Agarwal S, Johnston M, Cohen A. Are figure legends sufficient? evaluating the contribution of associated text to biomedical figure comprehension. J Biomed Discov Collab 2009;4(1).
2. Müller H, Kalpathy-Cramer J, Eggel I, Bedrick S, Radhouani S, Bakke B, et al. Overview of the CLEF 2009 medical image retrieval track; 2009.
3. Silvis ML, Clinch CR, Tillett JS, Spogen D. Clinical inquiries: What is the best way to evaluate an acute traumatic knee injury? J Fam Pract 2008;57(2):116–8.
4. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Methods Inf Med 1993;32(4):281–91.
5. Xu S, McCusker J, Krauthammer M. Yale Image Finder (YIF): a new search engine for retrieving biomedical images. Bioinformatics 2008;24(17):1968–70.
6. Kahn CE Jr, Thao C. Goldminer: a radiology image search engine. AJR Am J Roentgenol 2007;188(6):1475–8.
7. Hearst MA, Divoli A, Buturu H, Ksikes A, Nakov P, Wooldridge MA, et al. Biotext search engine: beyond abstract search. Bioinformatics 2007;23(16):2196–7.
8. Simpson M, Rahman MM, Demner-Fushman D, Antani S, Thoma GR. Text- and content-based approaches to image retrieval for the imageclef 2009 medical retrieval track; 2009.
9. Hernandez C, Ivkovic A, Fowler A. Growing plaque on foot. J Fam Pract 2008;57(9):603–5.
10. Kiss T, Strunk J. Unsupervised multilingual sentence boundary detection. Computational Linguistics 2006;32(4):485–525.
11. Yu H. Towards answering biological questions with experimental evidence: automatically identifying text that summarize image content in full-text articles. In: AMIA Annu Symp Proc. 2006, p. 834–8.
12. Ide NC, Loane RF, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc 2007;14(3):253–63.
13. Kahn CE Jr, Rubin DL. Automated semantic indexing of figure captions to improve radiology image retrieval. J Am Med Inform Assoc 2009;16(3):380–6.
14. Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics 2007;33(1):63–103.
15. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301–10.
16. Buckley C, Voorhees EM. Retrieval evaluation with incomplete information. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, United Kingdom: ACM; 2004, p. 25–32.
17. Smucker MD, Allan J, Carterette B. A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management. Lisbon, Portugal: ACM; 2007, p. 623–32.
18. Kalpathy-Cramer J, Hersh W. Automatic image modality based classification and annotation to improve medical image retrieval. Stud Health Technol Inform 2007;129(Pt 2):1334–8.