Reliability and Type of Consumer Health Documents on the World Wide Web: an Annotation Study

Melanie J. Martin
California State University, Stanislaus
One University Circle
Turlock, CA 95382
mmartin@cs.csustan.edu

Abstract

In this paper we present a detailed scheme for annotating medical web pages designed for health care consumers. The annotation is along two axes: first, by reliability or the extent to which the medical information on the page can be trusted, second, by the type of page (patient leaflet, commercial, link, medical article, testimonial, or support). We analyze inter-rater agreement among three judges for each category. Inter-rater agreement was moderate (0.77 accuracy, 0.62 F-measure, 0.49 Kappa) on the reliability axis and good (0.81 accuracy, 0.72 F-measure, 0.73 Kappa) along the type axis.

1 Introduction

With the explosive growth of the World Wide Web has come, not just an explosion of information, but also the explosion of false, misleading and unsupported information. At the same time, the web is increasingly used for tasks where information quality and reliability are vital, from legal and medical research by both professionals and lay people, to fact checking by journalists and research by government policy makers.

In particular, there has been a proliferation of web pages in the medical domain for health care consumers. At the first sign of illness or injury more and more people go to the web before consulting medical professionals. The quality and reliability of the information on consumer medical web pages has been of concern for some time to medical professionals and policy makers. (For example see Eysenbach et al., 2002, Impicciatore et al., 1997.)

Our goal is to create a system that can automatically measure the reliability of web pages in the medical domain (Martin, 2004). More specifically, given a web page resulting from a user query on a medical topic, we would like to automatically provide an estimate of the extent to which the information on the page can be trusted. In order to make use of supervised natural language processing and machine learning algorithms to create such a system, and to ultimately evaluate the performance of the system, it is necessary to have human annotated data.

It is important to note the varied uses of the term "reliability" in the computer and information sciences. In the current context we use it to refer to an intrinsic property of a web page: essentially the trustworthiness of the information it contains. This sense of reliability is distinct from its meaning in measurement theory as an indicator of repeatability. It also excludes measures such as credibility that are based on user beliefs or understanding.

In this paper we report results of an annotation study of medical web pages designed for health care consumers. Three humans annotated a corpus of web pages along two axes. The first axis is the reliability of the information contained in the page. The second axis is the type, or kind, of page. Intercoder agreement was moderate (0.77 accuracy, 0.62 F-measure, 0.49 Kappa) on the reliability axis and good (0.81 accuracy, 0.72 F-measure, 0.73 Kappa) along the type axis.

In our materials and methods section we discuss the data, definitions, annotation study and the results. We follow with a discussion section and a conclusion.

Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, pages 38–45,
Los Angeles, California, June 2010.

2 Materials and Methods

In this section we will discuss the data and definitions for the annotation task. We also describe the annotation study and the testing and analysis.

2.1 Data

The data to be annotated consists of two corpora of web pages created by the author: IBS70 and MMED100. The MMED100 corpus is a subset of a larger corpus (MMED1000). Both corpora are described below.

2.1.1 IBS70 Corpus

The IBS70 corpus was created as an exploratory corpus for use in system development. It was originally the top 50 Google hits for "irritable bowel syndrome" downloaded automatically though the Google API on July 1, 2004. The query was chosen to provide a range of quality and types of pages which one would expect to see more generally in the medical domain on the web: patient information from both traditional and alternative sources, support groups, medical articles, commercial pages from drug companies and quacks. During system development we determined that it would be useful to have additional pages at both ends of the reliability spectrum, possibly to use as seeds for clustering.

On September 15, 2004, twenty documents were added to the corpus to create the IBS70. Ten highly reliable documents were added based on web searches to find documents judged as meeting the standards of Evidence Based Medicine. Ten documents judged unreliable were added by taking the first ten relevant “Sponsored Links” resulting from a Google search on “irritable bowel syndrome”. There are two important things to note about this process: first, the high quality pages added were disproportionately from the U.K.; second, the low quality pages tend toward the crassly commercial and are more extreme than one would likely find in this proportion of the top 100 (or even 200) of the results of a Google query for a medical condition.

2.1.2 MMED100 Corpus

The MMED1000 corpus was created on November 5th and 8th, 2004 by automatically downloading from Google the top 100 search results for each of the following 10 queries:

- Adrenoleukodystrophy
- Alzheimer's
- Endometriosis
- Fibromyalgia
- Obesity
- Pancreatic cancer
- Colloidal Silver
- Irritable Bowel Syndrome
- Late Lyme Disease
- Lower Back Pain

The queries were chosen to provide a broad range of what might be typical queries for health consumers on the web and the types of pages that would result from these queries.

Colloidal Silver was chosen in the hopes of providing a sufficient number of pages of questionable reliability. Adrenoleukodystrophy, Pancreatic Cancer, Alzheimer’s and Obesity were chosen because there is general agreement in the medical community that these are diseases or health issues, and on diagnostic techniques. They also cover a spectrum of occurrence rates, with Adrenoleukodystrophy being relatively rare and Obesity being relatively common. The other five queries were chosen because there is less agreement in both the medical community and the general population about the existence, frequency, severity and treatment of these conditions. In particular, Fibromyalgia and IBS can be exclusionary diagnoses without clear and successful treatment options, which can open the door to web pages with a range of questionable treatments.

For annotation purposes a subset of this corpus, MMED100, with 100 pages, was created by randomly selecting ten documents from each of the ten queries.

At this time neither corpus is publicly available. However they can be provided on request and it is anticipated that they will be made publicly available once a viable standard is established for the annotations.

2.2 Definitions

The primary classification task is to classify pages based on their reliability (quality or trustworthiness of the information they contain). The secondary

classification task is to classify pages based on their type (e.g. commercial, patient leaflet, link). The classification by type emerged from the hypothesis that different types of pages may need to be treated differently to classify them based on their reliability. For example, if the primary purpose of a page is to provide links to information, determining the reliability of the page may require determining the reliability of the pages to which it links. However, in the current study, annotators are provided only the given web page and not allowed to follow links, so their reliability determination was made based on the apparent balance and objectivity of the links on the page.

For both tasks, only one tag was allowed, so annotators were instructed to consider the main purpose or intent of the page.

2.2.1 Reliability

Reliability of web pages is annotated based on a five level scale.

Probably Reliable (PrR)
The information on these pages appears to be complete and correct, meeting the standards of Evidence-Based Medicine where appropriate. Information is presented in a balanced and objective manner, with the full range of options discussed (where appropriate). The page and author appear reputable, with no obvious conflicts of interest. The appropriate disclaimers, policies, and contact information are present. Where appropriate, sources are cited. An example of a page in this category would be a patient leaflet from a reputable source that adheres to the standards of Evidence-Based Medicine.

Possibly Reliable (PoR)
The information on the page is generally good and without obvious false or outdated statements, but may not be sufficiently complete and balanced or may not conform to evidence-based standards. An example of a page in this category would be a patient leaflet that contains only a brief description of diagnostic procedures or suggests a treatment option that is generally accepted, but not supported by evidence.

Unable to determine (N)
For these pages it is difficult or impossible to determine the reliability, generally because there is not enough information. For example, the page may be blank, only contain login information, or be the front page of a medical journal.

Possibly Unreliable (PoU)
These pages may contain some reliable information, but either have some that is outdated, false or misleading, or the information is sufficiently unbalanced so as to be somewhat misleading. An example of a page that might fall into this category is a practitioner commercial pages, which has valid information about an illness, but only discuss the preferred treatment offered by the practitioner.

Probably Unreliable (PrU)
These pages contain false or misleading information, or present an unbalanced or biased viewpoint on the topic. Examples of pages in this category would include: testimonials (unsupported viewpoints or opinions of a single individual) or pages that are clearly promoting and selling a single treatment option.

2.2.2 Type of Page

We found six types of pages that frequently come up in search results for queries in the medical domain: Commercial, Patient Leaflet, Link, Medical Articles, Support, and Testimonials. There are also pages which are not relevant, or do not contain sufficient information to make a determination. Below we discuss each of these types. When a page seems to overlap categories the annotation is based on the primary purpose of the page.

Commercial (C)
The primary purpose of these pages is to sell something, for example, pages about an ailment sponsored by a drug (also more general treatment or equipment) company, which sells a drug to treat it. Given the desire to sell, these pages might not present complete or balanced information (making them less likely to be reliable). Practitioner pages with no real (substantial) information, which are designed to get people to make an appointment, as opposed to patient leaflets (designed to supplement information that patients receive in the office or clinic), might also fall into this category

Link (L)
The primary purpose of these pages is to provide links to other pages or sites (external), which will provide information about a certain illness or medical condition. These links may or may not be annotated, and the degree of annotation may vary considerably. Since the reliability of these pages depends on the reliability of the pages they link to (possibly also on the text in the annotations), without following the links a reliability estimate can be based on the range and apparent objectivity of the links.

Patient Leaflet, Brochure, Fact Sheet or FAQ (P)
The primary purpose of these pages is to provide information to patients about a specific illness or medical condition. Generally, these pages will be produced by a clinic, medical center, physician, or government agency, etc. The primary purpose is to provide information. This class needs to be distinguished from medical articles, especially in encyclopedias or the Merck Manual, etc. These pages will tend to have headings like: symptoms, diagnosis, treatment, etc. These headings can take the form of links to specific parts of the same page or to other pages on the same site (internal). The reliability of these pages is based on their content and determined by factors including Evidence-Based Medicine, completeness, and the presence of incorrect or outdated information.

Medical Article (practitioner or consumer) (MA)
The primary purpose of these pages is to discuss an aspect of a specific illness or medical condition, or a specific illness or medical condition. These can be divided into two main categories: articles aimed at consumers and articles aimed at health practitioners.

Articles aimed at health practitioners, particularly doctors, may be scientific research articles. The reliability of these pages is based on their content and determined by factors including Evidence Based Medicine, completeness, and the presence of incorrect or outdated information. Note: Medline search results may be considered a links page to medical articles.

Articles aimed at consumers may come from a variety of sources including mainstream and alternative media sources. Reliability is determined based on the content as with articles for practitioners.

Testimonial (T)
The primary purpose of these pages is to provide testimonial(s) of individuals about their experience with an illness, condition, or treatment. While individuals may be considered reliable when discussing their own personal experiences, these pages tend to be unreliable, because they are generally not objective or balanced. There is a tendency for readers to generalize from very specific information or experiences provided by the testimonial, which can be misleading.

Support (S)
The primary purpose of these pages is to provide support of sufferers (or their loved ones or caregivers) of a particular illness or condition. The pages may contain information, similar to that found in a patient leaflet; links to other sites, similar to a links page; and testimonials. In addition they may contain facilities such as chat rooms, newsletters, and email lists. Activities may include lobbying for funding for research, generally put up by individuals or non-profit organizations. For reliability, one may need to look at the agenda of the authors or group. It may be in their interest (politically) to overstate the problem or make things out to be worse then they are to secure increased funding or sympathy for their cause.

Not Relevant (N)
These pages are blank or not relevant and include: login pages, conditions of use pages, and medical journal front pages.

2.3 Annotation Study

In order to get started with system development, a single annotator, M, who was involved with development of both the classifications and the system, tagged the IBS70 and MMED100. Then in Spring 2008 two senior undergraduate science majors (chemistry and biology), L and E, were hired for the annotation study. The annotation study consisted to two primary phases: training and testing. Each phase is described below.

2.3.1 Training Phase

The two student annotators, L and E, received copies of the draft annotation instructions. They each met individually with M to discuss the instructions and any questions they had.

For each of three training runs, ten randomly chosen web pages from the IBS70 corpus were posted on a private web site. The students annotated the pages for reliability and type and then met individually to discuss their annotations with M. As questions and issues arose, the instructions were amended to reflect clarifications. For example, L needed additional instructions on the distinction between Link and Patient Leaflet pages; a separate category for FAQs was collapsed into the Patient Leaflet category.

2.3.2 Testing Phase

Once the student annotators seemed to be achieving reasonable levels of agreement (Cohen's Kappa above 0.4) on each task, there was a three-part testing phase. The remaining 40 pages in the IBS70 corpus were randomly divided into two test corpora and finally the MMED100 corpus was annotated.

During the testing phase, one of the students, L, seemed to annotate less carefully. (Possibly because the timing coincided with graduation and summer vacation.) For example, on the MMED100 corpus L tagged 30% as N (unable to determine the reliability, compared to 12% for E and 10% for M. L was asked to go back and reconsider the web pages tagged as N. We report results with L's reconsidered tags here for completeness, but further discussion will focus on agreement between M and E.

2.4 Testing and Analysis

We report inter-rater agreement using accuracy, Cohen's Kappa statistic (Cohen, 1960) for chance corrected agreement and F-Measure (Hripcsak and Rothschild, 2005). We consider each annotation axis separately.

2.4.1 Page Reliability

We can estimate a baseline distribution of the categories R (reliable), N (unable to determine), and U (unreliable) based on an average of the tags across all training and test sets to be approximately: 68% R; 13% N; 19% U.

Table 1 shows the results for the Accuracy (percent agreement) and Kappa statistic on the five reliability classes across all the corpora. It became immediately clear the annotators were not able to make the more fine-grained distinctions between "probably" and "possibly" for either the reliable or unreliable classes, given the current instructions and timeline. The classes were then collapsed to three: R (reliable), N (unable to determine) and U (unreliable) and the results are shown in Table 2.

Accuracy/ Kappa	5 Classes Reliability		
Set\Raters	M-E	M-L	E-L
IBS train	**0.47**/0.30	**0.33**/0.12	**0.40**/0.19
IBS test	**0.33**/0.11	**0.40**/0.25	**0.43**/0.28
MMed100	**0.51**/0.32	**0.35**/0.12	**0.38**/0.14

Table 1. Inter-rater agreement for 5-class reliability.

Accuracy/ Kappa	3 Classes Reliability		
Set\Raters	M-E	M-L	E-L
IBS train	**0.70**/0.44	**0.60**/0.25	**0.67**/0.33
IBS test	**0.70**/0.43	**0.65**/0.42	**0.75**/0.59
MMed100	**0.77**/0.49	**0.66**/0.30	**0.62**/0.22

Table 2. Inter-rater agreement for 3-classs reliability.

The results in Table 2 for M-E show improved agreement after training and consistent moderate agreement on the test corpora based on the Kappa statistic. Accuracy (percent agreement) for M-E is 70% for both IBS testing and training and 77% for the MMED100.

Further analysis of L's reliability tags showed a bias toward the "U" tag. For example, in the MMED100 corpus, L tagged 28% as U, compared to 19% and 17% for M and E, respectively.

Hripcsak and Rothschild (2005) suggest use of the F-measure (harmonic average of precision – equivalent to positive predictive value - and recall – equivalent to sensitivity - commonly used in Information Retrieval) to calculate inter-rater agreement in the absence of a gold standard. In Table 3 we report the average F-measure between each pair of raters and the F-measure by class. A higher F-measure indicates better agreement, so these results show that the "Can't Tell" class is the most

difficult to agree on, followed by the "Unreliable" class.

MMED100 F-Measure	3 Classes Reliability		
Class\Raters	M-E	M-L	E-L
Reliable	0.87	0.78	0.76
Can't Tell	0.45	0.22	0.30
Unreliable	0.55	0.46	0.36
Average	**0.62**	**0.49**	**0.47**

Table 3. F-measure by class for 3-classs reliability.

In order to look for patterns of agreement between the raters we looked at agreement by query in the MMED100 corpus. In Table 4 we show the agreement for M and E by query. Although it appears that some queries were easier to annotate than others, since there are only 10 pages per query, the sample may be too small to draw definite conclusions.

Query	Accuracy	Kappa
Endometriosis	1	1
Pancreatic Cancer	1	1
Late Lyme	1	1
Adrenoleukodystrophy	0.8	0.412
Obesity	0.8	0.655
Alzheimer's	0.7	-0.154
Fibromyalgia	0.7	0.444
Lower Back Pain	0.7	-0.154
Colloidal Silver	0.6	0.13
Irritable Bowel Syndrome	0.4	-0.053

Table 4. Inter-rater reliability agreement for M-E by query.

Possible ways to improve these results are presented in the "Discussion" section.

2.4.2 Page Type

The dominant page types are P (patient leaflets), L (link), C (commercial) and MA (medical article). The baseline distribution based on averages across the training and test sets is approximately: 39% P; 15% L; 18% C; and 13% MA. The other three classes S (support), T (testimonial), and N (unable to determine) making up only 15% of the pages in the corpus.

Table 5 shows the results for Accuracy and the Kappa statistic on the seven type classes across all the corpora. Collapsing categories for the type annotation task did not appreciably increase Kappa scores (M-E Kappa was 0.742 on the MMED100 corpus when the P and MA classes were collapsed), so it seems preferable to keep the original classes.

Accuracy/ Kappa	Type		
Set\Raters	M-E	M-L	E-L
IBS train	**0.57**/0.42	**0.83**/0.78	**0.47**/0.28
IBS test	**0.73**/0.64	**0.65**/0.55	**0.73**/0.64
MMed100	**0.81**/0.73	**0.48**/0.29	**0.50**/0.31

Table 5. Inter-rater agreement for type annotation.

Again we see with annotators M and E, the improved agreement from training to testing, as distinctions between classes were clarified (for example, between Link and Patient Leaflets, and between Patient Leaflets and Medical Articles).

We also computed F-measure by type for the MMED100 corpus, as shown in Table 6. Of the three most common types of pages (Patient Leaflet, Link, Commercial), the Link type was the most difficult for M-E to agree on.

MMED100 F-Measure	Type		
Class\Raters	M-E	M-L	E-L
P	0.893	0.593	0.625
L	0.625	0.480	0.435
C	0.727	0.323	0.414
S	0.769	0.222	0.250
T	0.500	0.000	0.800
MA	0.667	0.593	0.455
N	0.857	0.143	0.118
Average	**0.720**	**0.336**	**0.442**

Table 6. F-measure by class for page type.

We further analyzed the page type annotations by query for raters M and E (Table 7). We found a negative correlation between the variance of the types in a query to the Kappa statistic of agreement for the query ($r^2 = -0.62$).

Query	Accuracy	Kappa
Endometriosis	0.9	0.851
Fibromyalgia	0.9	0.846
Alzheimer's	0.8	0.75
Irritable Bowel Syndrome	0.8	0.73
Obesity	0.8	0.697
Pancreatic Cancer	0.8	0.63
Colloidal Silver	0.8	0.63
Adrenoleukodystrophy	0.8	0.512
Lower Back Pain	0.8	0.512
Late Lyme	0.7	0.483

Table 7. Inter-rater type agreement for M-E by query.

3 Discussion

Librarians, scholars, and information scientists have done significant work on the quality (reliability) of print, and more recently, web information (for example, see Cook 2001, Alexander and Tate 1999). It is important to distinguish quality (reliability) from credibility (e.g. Danielson 2005), which is based on the users view of the information. Here we are interested in the quality of the information itself.

In a relatively early study, Impicciatore et al. (1997) sampled web documents relating to fever in children and found the quality of the information provided to be very low. In 2002, Eysenbach et al. conducted a review of studies assessing the quality of consumer health information on the web. Of the 79 studies meeting their inclusion criteria (essentially appropriate scope and quantitative analysis), they found that 70% of the studies concluded that reliability of medical information on the Web is a problem.

To address the question of how to determine the quality of medical information on the web, Fallis and Frické (2002) empirically tested several proposed indicators and found that the standard indicators of quality for print media could not be directly translated to consumer medical information on the Web. Price and Hersh (1999) developed a semi-automated system to filter out low quality consumer medical web pages based on approximately 30 criteria.

Annotation studies have been discussed and conducted in the computational linguistics community for a variety of annotation tasks, including subjectivity (e.g. Weibe et al. 1999) and opinion (e.g. Somasundaran et al. 2008). Artstein and Poesio (2008) surveyed inter-coder agreement in computational linguistics, including Cohen's Kappa.

To ensure a "gold standard" for training machine learning algorithms to do automatic classification a number of approaches could be pursued: the production of bias-corrected tags as described by Weibe et al. (1999); a new study with "expert" annotators – having a stronger medical background – and additional training; ask annotators to use existing web tools (e.g. American Accreditation HealthCare Commission) to assess the page quality; systematically assess whether the noise introduced by moderate agreement levels will create problems for machine learning with this data (Beigman Klebanov and Beigman 2009).

The agreement on the type annotation task could still be improved, possibly by additional clarification to the definitions. However, it is still to be determined if noise levels are low enough and sufficiently random to be used successfully in supervised learning. This task is easier than the reliability task and requires less expertise of the annotators.

4 Conclusion

There is a demonstrated need to provide tools to health care consumers to automatically filter web pages by the reliability, quality, or trustworthiness of the medical information the pages contain. We have shown promising results in this study that appropriate classes of pages can be developed. These classes can be used by human annotators to annotate web pages with reasonable to good agreement.

Thus we have laid a foundation for future annotation studies to create a gold standard data set of consumer medical web pages. The corpora in this study are currently being used to create an automated system to estimate the reliability of medical web pages.

Acknowledgments

This work was supported in part by a CSU Stanislaus Naraghi Faculty Research Enhancement Grant. I am grateful to Elizabeth Jimenez and Luis Adalco for participating in the annotation study and to the anonymous reviews for their comments and suggestions. I would also like to thank Roger Hartley my dissertation advisor and Peter Foltz for discussions during the formulation and develop-

ment of the system, and Tom Carter for helpful and insightful comments leading to the improvement of this paper.

References

Janet E. Alexander and Marsha Ann Tate. 1999. *Web Wisdom: How to Evaluate and Create Information Quality on the Web.* Lawrence Erlbaum and Associates, New Jersey.

American Accreditation HealthCare Commission. *Health information on the internet: A checklist to help you judge which websites to trust.* Retrieved February 28, 2010, from http://www.urac.org

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.* 34, 4 (Dec. 2008), 555-596.

B. Beigman Klebanov, and E. Beigman. 2009. From annotator agreement to noise models. *Comput. Linguist.* 35, 4 (Dec. 2009), 495-503.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, pages 34-46.

Alison Cooke. 2001. *A Guide to Finding Quality Information on the Internet: Selection and Evaluation Strategies, Second Edition.* Library Association Publishing, London.

D.R. Danielson. 2005. Web credibility. C. Ghaoui (Ed.), *Encyclopedia of Human-Computer Interaction.* Hershey, PA: Idea Group, 713-721.

Gunther Eysenbach, John Powell, Oliver Kuss, and Eun-Ryoung Sa. 2002. Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web: A Systematic Review. *JAMA*, May 22, 2002; 287(20): 2691 - 2700.

Don Fallis and Martin Frické. 2002. Indicators of Accuracy of Consumer Health Information on the Internet. *Journal of the American Medical Informatics Association,* 9, 1, (2002): 73-79.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the F-Measure, and Reliability in Information Retrieval. *J Am Med Inform Assoc*. 2005 May–Jun; 12(3): 296–298.

Piero Impicciatore, Chiara Pandolfini, Nicola Casella, and Maurizio Bonati. 1997. Reliability of Health Information for the Public on the World Wide Web: Systematic Survey of Advice on Managing Fever in Children at Home. *BMJ* 1997; 314:1875 (28 June).

Melanie J. Martin. 2004. Reliability and Verification of Natural Language Text on the World Wide Web. Paper at *ACM-SIGIR Doctoral Consortium*, July 25, 2004, Sheffield, England.

Susan L. Price and William R. Hersh. 1999. Filtering Web Pages for Quality Indicators: An Empirical Approach to Finding High Quality Consumer Health Information. *American Medical Informatics Association* 1999.

Swapna Somasundaran, Josef Ruppenhofer and Janyce Wiebe. 2008. Discourse Level Opinion Relations: An Annotation Study. Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue Columbus, Ohio, June 19-20, 2008, pp. 129–137.

Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association For Computational Linguistics on Computational Linguistics* (College Park, Maryland, June 20 - 26, 1999). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 246-253.