Using Shallow Semantic Analysis to Implement Automated Quality Assessment of Web Health Care Information

Yanjun Zhang[1], Robert E. Mercer[1], Jacquelyn Burkell[1], and Hong Cui[2]

[1] The University of Western Ontario, London, Ontario, Canada
[2] University of Arizona, Tucson, Arizona, USA

Abstract. Evidence-based clinical practice guidelines have been widely used as an objective rating instrument for assessing the content quality of health care information on the web. In many previous studies, human raters check the concordance between text content and evidence-based practice guidelines in order to evaluate information accuracy and completeness. However, human rating cannot be a practical solution, particularly when there is an extremely large volume of health care information on the web. This study explores a semantics-based approach to identify health care information content in web documents with reference to evidence-based health care guidelines. With this approach terms and phrases in English are extracted and transformed into semantic concepts and units. Thus, web text is transformed, sentence by sentence, into a semantic representation which computer programs can classify depending on whether the content of a sentence is in concordance with evidence-based guidelines or not. Through aggregating the classification result of all sentences in a web document, computer programs are able to generate for each document a quality score indicating the number of unique evidence-based guidelines that are referred to in the document. In a test using a set of depression treatment web pages and evidence-based clinical guidelines, the quality rating performance of the computer system is shown to be close to human quality rating performance.

Keywords: Semantics-based quality rating approach, natural language processing, shallow semantic analysis, evidence-based clinical practice, semantics-based classification, web health care information

1 Introduction

The past decade has witnessed a dramatic expansion in the amount of publicly available health care information on the Web. On one hand, the demand for web health care information is huge and keeps growing (Pew Internet and American Life Project, 2003; 2011; Podichetty et al., 2006). On the other hand, the information quality is of extreme variability (Eysenbach et al. 2002; Kunst et al. 2002; Griffiths & Christensen, 2005). In spite of this, different previous studies identified that users are ready to accept health care information from unfamiliar websites and they do not like the burden to verify information quality (Pew Internet and American Life Project, 2006). Such situations can cause threat to public health, including life-threatening cases (Crocco et al, 2002; Kiley, 2002). Because of the potential harm that may be caused by inaccurate information, quality assessment of health care information on the web stays a common interest of various health care information stakeholders, including e-health policy makers, information providers/consumers, and information search service providers.

In the last ten years and even longer, researchers have made a lot of efforts and progress in the quality evaluation area, including exploration on establishing quality rating criteria, creating rating tools, etc. One type of commonly used rating criteria is evidence-based health care practice guidelines. As summarized in a systematic review (Eysenbach et al., 2002) based on 79 distinct quality evaluation studies, health care guidelines are widely utilized by researchers for rating content quality in terms of information accuracy and comprehensiveness. Evaluating such type of quality is just the focus of this study. Our goal is to develop an automated approach to evaluate the information accuracy and comprehensiveness of health care web pages. Other types of quality, such as web site/page design aesthetics, web page readability, etc., referred as presentation quality in (Eysenbach et al., 2002) are not covered in this study. Depression treatment is the selected subject domain.

2 Related Work

Based on summarization from 79 distinct quality evaluation studies, Eysenbach et al. (2002) defined accuracy as the degree of concordance between the web content and generally accepted health care practice. The completeness is also called "comprehensiveness" or "coverage". Most researchers calculated the proportion of pre-defined clinical guidelines covered by a web source. In practice, the evaluation of accuracy and comprehensiveness requires human raters to read and understand web page content in order to compare web content against guidelines. The objectiveness and effectiveness of guidelines are advantages compared with some other criteria such as accountability meta-information of web sites or pages (Burkell, 2004). However, the use of health care guidelines for rating health care web information quality is limited by the dependency on human efforts. Due to the explosion of health care information on the web, it would be an impractical practice to have human raters manually evaluate all related web pages through reading content sentence by sentence.

On the other hand, many previous studies have investigated the association between content quality and accountability of web sites or pages, and attempted to use the latter as quality indicators since they are subject independent. Investigated indicators range from bibliographic metadata (e.g. authorship, editorial board, references) from the print world (e.g. Silberg et al., 1997, Chen et al., 2000; Smith, 2002; Barnes et al., 2009) to web-unique features (e.g. Frické et al., 2005; Griffiths & Christensen, 2005) such as site domain name suffix, hyperlinks received from external web sites, Google PageRank, etc., and to quality seal such as HONcode certificate (HONcode, 2012). Specifically, Wang and Liu (2007) developed an Automatic Indicator Detection Tool to collect indirect quality indicators using an HTML parser. In their testing, the performance of detecting such indicators reached 93% recall and 98% precision. However, their study did not include quality evaluation test to prove that detected indicators can accurately predicate content quality. In fact, different research groups (Frické et al., 2002, 2005; Martin-Facklam et al. 2003; Griffiths et al., 2005; Khazaal et al., 2012) found in their studies that the association between these indicators and the content quality of web health care information is inconsistent in different health care subjects, putting the validity and reliability of these non-content based indicators in question.

In comparison, rating the quality directly based on the web site/page content is relatively more reliable than relying on metadata indicators. To the best of our knowledge, Griffiths et al. (2005) did a first attempt of automated quality rating based on processing the web text content. In their study, they used evidence-based depression treatment guidelines (CEBMH, 1998) as rating criteria, and human rating quality scores according to guidelines are standards for evaluating automated rating performance. Their approach is based on information retrieval techniques. They tried to use web pages from training web sites to establish two standard queries, comprising of 20 keywords and 20 two-word phrases with high discriminative power in terms of content quality or relevance respectively. For given web sites, the similarity between the web content and the standard queries are used to calculate the site quality score. In their testing, the automatically rated website scores and the evidence-based human rated scores have strong Pearson correlation equal to 0.85.

3 Method

Griffiths et al. (2005) successfully used the keywords among health care web pages to predicate the information quality. In our study, we try to utilize the semantics of web content, and specifically analyze semantics at sentence level in order to predicate whether a sentence present meanings in concordance with given health care guidelines. Two key issues are explored and solved in order to reach this goal:

1. First, how to create an effective representation of the web text semantics? To compare the web content against evidence based health care guidelines, it is important that our automated quality rating approach can capture and represent text semantics at appropriate extent.
2. Second, with captured semantics, how can computer successfully identify the presentation of health care guidelines in web document?

Semantics-based quality rating approach: overall, this automated rating approach use computer programs to read in every sentence in a web page and check if its content agrees with any pre-defined health care guideline. The quality score is the number of unique guidelines that are referred to in the web page.

Text classification is applied to a single sentence to determine if it presents one or more guidelines. Each guideline has a binary classifier to categorize sentences into “positive” (i.e. a match with guideline) or “negative” group. Features used to implement text classification mainly include text semantics and relevant metadata. Shallow semantic analysis is used to capture these features and map a sentence into a semantic tag instance. Thus, text classification is done based on semantic tags rather than the original text.

Shallow Semantic Analysis: as the purpose is to convert sentences into semantic tag instances, shallow semantic analysis in this study focuses only on annotating semantically essential units in a sentence. Certainly, every depression treatment guideline has a unique theme. We assume that it contains a specific set of semantic concepts although they could have different variations when expressed in natural language.

First, health care concepts such as health conditions, treatment etc. are one type of the very important semantic units. They are usually nouns or noun phrases. In addition, verbs, adjectives and adverbs are important for describing the relations between semantic concepts. We want to map these elements from text to semantic tags, and also get their part of speech, and their distance between each other. We consider distance as a useful feature because semantic unit pairs which have tight relations likely occur close to each other.

To capture the semantic concepts and above features, we used the Unified Medical Language System (UMLS, 2009) to develop our shallow semantic tagging application. UMLS is a free resource provided by the National Library of Medicine in the United States. It provides a knowledge source called Metathesaurus, which include more than 60 controlled vocabularies in biomedical domains such as MeSH, SNOMED, etc.

UMLS also provides supporting software tools to facilitate access and use of UMLS data. Two very useful software packages are used in this study: 1) MetaMap API is used to discover Metathesaurus concepts referred by text. The benefit is that health care terms and variants of a same semantic concept can be unified. For example, text strings “depression”, “Depressive episode” and “depressive illness” are all labeled into MMTx tag “Depressive disorder”. 2) SPECIALIST NLP Tools (including LVG and TaggerClient) facilitate natural language processing by dealing with lexical variation and text analysis tasks in the biomedical domain. Supported NLP functions include sentence splitting, tokenization, POS tagging, lemmatization, etc. We used this tool to transform and filter lexical variants from the original text of sentences. For example, “ceases”, “ceased”, “stopping”, and “stops” can be transformed to a LVG tag “stop”.

We developed JAVA programs to utilize above resources and to use the UMLS tagging results to generate semantic tag instance for sentences. The definition of a semantic tag instance includes header and body parts (see Figure 1.). The instance header includes sentence sequence number in a page, begin and end offset of this sentence, and the original text. In the instance body, the labeled semantic units are saved in sequence. Each labeled unit is either a LVG tag by default or an MMTx tag for Metathesaurus concepts, enclosed by a pair of square brackets and separated by pipelines.

Syntax of Semantic Tag Instance (for a sentence):
|===sentence===|sentenceSequenceNum|begin-offset|end-offset|#%#original text of sentence%#% [LVG or MMTx tag]|[LVG or MMTx tag]|. . .

LVG|MMTx Tag Syntax:
[tag, begin-offset, end-offset, Term Sequence Num within the hosting sentence, POS|Semantic Type, (Semantic mapping score)]

Fig. 1. Definition of semantic tag instance

Figure 2 shows two examples of semantic tagging result. A semantic tag instance contains tags for semantically essential words or phrases in a sentence. In the first example, phrase “depressive illness” is converted to a unified concept “Depressive Disorder”. In the second, “tricyclics” was converted to concept “Antidepressive Agents”. In addition, each tag also contains the position metadata, POS metadata, etc. For example, “effect side” is the semantic tag for text “side effects” which has position index from 365 to 376. The term index of this semantic unit in this sentence is 0. The part of speech is NOUN.

1. People with a depressive illness cannot merely "pull themselves together" and get better.
|===Sentence===|8|520|610|#%#People with a depressive illness cannot merely "pull themselves together" and get better. %#%[Persons,520,525,0,Population Group,1000]| [Depressive disorder, 534,543,3, Mental or Behavioral Dysfunction,1000]|[merely,560, 565,6, adv]|[pull,569,572,7,adv] |[together,585,592,9,adv]|[get,599,601,11,verb]|[best,603,608,12,adj]
2. Side effects of tricyclics, which vary from person to person, may include dry mouth, blurred vision, constipation, problems passing urine, sweating, light-headedness and excessive drowsiness.
|===Sentence===|3|365|557|#%#Side effects of tricyclics, which vary from person to person, may include dry mouth, blurred vision, constipation, problems passing urine, sweating, light-headedness and excessive drowsiness.%#%[effect side,365,376,0,noun] |[Antidepressive Agents,381,390,2,Organic Chemical, Pharmacologic substance,1000]|[vary,399, 402,4,verb] |[Persons,409, 414,6,Population Group,1000]|. . . |[excessive,535,543,20,adj]|[Drowsiness,545, 554,21,Sign or Symptom,861]

Fig. 2. Example of semantic tag instances

Rule-based Classification: The generated semantic tag instances are the input for text classification. Each guideline has a dedicated classifier to make binary prediction regarding whether a given sentence agrees with the specific guideline. This study used a rule-based classification to solve the problem. For each guideline, rules (i.e. classification patterns) were manually summarized based on studying the positive instances in training data set.

In this study, a classification rule is considered as a description of relations between semantic units which are indispensable for identifying the presentation of a specific health care guideline. Considering the variation of natural language expression, we believe that a classifier can have multiple classification rules and each corresponds to a distinct expression pattern. Such patterns are extractable from positive sentences. The knowledge engineering for extracting patterns starts from identifying the theme-related semantic units, then try to find connections between them. For most patterns, usually more than one positive instance can be found from training data. Features that commonly exist in the positive instances are used to define the classification pattern. The features that were used in this study include semantic tag, co-occurrence, part-of-speech, distance between key tags, tag ordering relations, and negative proposition.

Figure 3 lists out the XML-formatted classification rules for guideline #6. It has 3 patterns that were identified from positive training instances. Each pattern is comprised of multiple semantic units. For example, the content listed between the pair of XML markup <Pattern> and </Pattern> describes the first classification pattern. Key semantic units in this pattern include antidepressant, side effect, vary and so on. These required semantic units are called patternUnit. The identity of a patternUnit is defined using LVG or MMTx tag. Under each patternUnit, constraints in terms of distance between a pair of patternUnits and the sequence of their occurrences are defined to describe the co-occurrence relationship between patternUnits, whenever such relationship exists. In order to classify a semantic tag instance, which represents a sentence, into TRUE, the specified constraints need to be satisfied. The example here is that the semantic unit "vary" needs to co-occur with another unit, "side effect", either BEFORE or AFTER, with interval terms or phrases no more than 5. These values are configured based on statistics from training instances.

The classification working logic is simple. It reads in the semantic tag instance of a sentence; then matches it against the pre-defined classification patterns for a specific health care guideline. During the matching, computer scan the semantic units within a semantic tag instance to search for patternUnits required by the classification and calculate the matching probability based on the searching result. If any fitting pattern is confirmed, then classification result is TRUE, otherwise FALSE.

Quality Scores: Given a web page, its quality rating score is automatically generated based on the sentence classification results. The classifiers classify the test instance (i.e. sentence) as either Positive or Negative regarding to specific guidelines. If it is positive, the webpage containing the sentence scores one. However, the presentation of a same guideline in one web page will be counted only once. This complies with the standard used human rating. So, the final quality score is equal to the number of unique guidelines that rule-based classifiers identified in a web page.

```
<RulePattern>
<ruleID>6</ruleID>
<patAmount>3</patAmount>
<!– "antidepressant", "side effect",
"vary", "not"(NEGPunit), proxim-
ity(2,3)=[EITHER,5] –> <Pattern>
<PID>1</PID>
<punitAmount>4</punitAmount>
<punit>
<eID>1</eID>
<keyword>Antidepressive
Agents</keyword>
<tagType>MMTx-1</tagType>
<pos>N</pos>
<synset>
<synCount>3</synCount>
<syn>
<term>MAOIs?</term>
<tagType>TEXT</tagType>
<pos>N</pos>
</syn>
<syn>
<term>SSRIs?</term>
<tagType>TEXT</tagType>
<pos>N</pos>
</syn>
<syn>
<term>SNRIs?</term>
<tagType>TEXT</tagType>
<pos>N</pos> </syn>
</synset>
<term>SSRIs?</term>
<tagType>TEXT</tagType>
<pos>N</pos>
<term>SNRIs?</term>
<tagType>TEXT</tagType>
<pos>N</pos>
<alter_in_context>
<altCount>2</altCount>
<alternative>
<term>Pharmaceutical
Preparations</term>
<tagType>Hypernym</tagType>
<pos>N</pos>
</alternative>
<!– This MMTx includes free text Medi-
cation, medicine and drug –>
<alternative>
<term>drug</term>
<tagType>Hypernym</tagType>
<pos>N</pos>
</alternative>
</alter_in_context>
<enforce>1</enforce>
<co-occurrence>
<co-flag>N</co-flag>
</co-occurrence>
</punit>
<punit>
<eID>2</eID>
<keyword>effect side</keyword>
<tagType>LVG</tagType>
<pos>N</pos>
<synset> <synCount>1</synCount>
<syn>
<term>side-?effects?</term>
<tagType>TEXT</tagType>
<pos>unknown</pos>
</syn>
</synset>
<alter_in_context>
<altCount>0</altCount>
</alter_in_context>
<enforce>1</enforce>
<co-occurrence>
<co-flag>Y</co-flag>
<cotermContainer>
N
</cotermContainer>
</co-occurrence>
</punit>
<punit>
<eID>3</eID>
<keyword>vary</keyword>
<tagType>LVG</tagType>
<pos>V</pos>
<synset>
<synCount>2</synCount>
<syn>
<term>change</term>
<tagType>LVG</tagType>
<pos>V</pos>
</syn> <syn>
</syn> </synset>
<term>alter</term>
<tagType>LVG</tagType>
<pos>V</pos>
<alter_in_context>
<altCount>1</altCount>
<alternative>
<term>differ</term>
<tagType>LVG</tagType>
<pos>V</pos>
</alternative>
</alter_in_context>
<enforce>1</enforce>
<co-occurrence>
<co-flag>Y</co-flag>
<cotermContainer>
Y
</cotermContainer>
<co-term>
<co-eid>2</co-eid>
<co-occur_proximity>
5
</co-occur_proximity>
<position_relation>
EITHER
</position_relation>
</co-term>
<!– one PUNIT is allowed to have mul-
tiple co-occurring PUNITs–>
</co-occurrence>
</punit>
<punit>
<eID>4</eID>
<keyword>not</keyword>
<tagType>LVG</tagType>
<pos>ADV</pos>
<synset>
<synCount>3</synCount>
<syn>
<term>never</term>
<tagType>LVG</tagType>
<pos>ADV</pos>
</syn> <syn>
</syn> <syn>
</syn> </synset>
<term>no</term>
<tagType>LVG</tagType>
<pos>ADJ</pos>
<term>no</term>
<tagType>TEXT</tagType>
<pos>unknown</pos>
<alter_in_context>
<altCount>3</altCount>
<alternative>
<term>unlikely</term>
<tagType>LVG</tagType>
<pos>N</pos>
</alternative>
<alternative>
<term>barely</term>
<tagType>LVG</tagType>
<pos>N</pos>
</alternative>
<alternative>
<term>rarely</term>
<tagType>LVG</tagType>
<pos>N</pos>
</alternative>
</alter_in_context>
<enforce>-1</enforce>
<co-occurrence>
<co-flag>Y</co-flag>
<cotermContainer>
Y
</cotermContainer>
<co-term>
<co-eid>3</co-eid>
<co-occur_proximity>
4
</co-occur_proximity>
<position_relation>
BEFORE
</position_relation>
</co-term>
</co-occurrence>
</punit> </Pattern>
<Pattern> ...
</Pattern> </RulePattern>
```

Fig. 3. Example of classification rules for health care guideline #6

4 Quality Rating Test

Data: This study testifies the semantics-based automated quality rating approach using depression treatment web pages. The whole corpus includes 201 depression treatment web pages (see author's thesis) was collected in 2009 from three types of sources as listed in Table 1. For search engines, "depression treatment" was used as the query and the first 30 returned web pages from each search engine were collected as candidates. For web portals, candidate pages were collected from depression treatment related sections only. Candidate pages were examined manually to remove duplicate pages and pages that were inappropriate for other reasons. In the end, 201 web pages were selected to form the corpus.

Table 1. Sources for Constructing the Corpus

Web Search Engines	Medical Search Engines	Health Care Web Portals
Google Yahoo! Search Microsoft Bing Search Ask.com HealthFinder	AOL OmniMedicalSearch HealthFinder HealthLine MedNar WebMD	Medline Plus in United States, HealthlinkBC in Canada, HealthInsite in Australia, National Health Service (NHS) in United Kingdom

All 201 pages were rated by human raters with reference to a set of evidence-based depression treatment guidelines (see author's thesis) previously used in (Griffith & Christensen, 2005; Griffiths et al., 2005). Human rated quality scores for all these pages range from 0 to 8. The pages were divided into five bins according to scores, i.e. 0, 1-2, 3-4, 5-6, and 7-8. Stratified random sampling was conducted to get 31 pages as testing data set, and the remaining 170 pages formed training data.

Measures for Classification Performance: The testing data set has 31 web pages and in total 2677 sentences. Precision, recall and accuracy were used to evaluate the sentence classification performance. The measures are calculated using the following equations: Precision= TP / (TP + FP), Recall= TP / (TP + FN), Accuracy= (FP + TN) / (TP + FN + FP + TN). TP stands for the number of true positive (cases) identified by classifier; FP stands for false positives identified by classifier; FN stands for false negatives identified by classifier; and TN stands for true negatives identified by classifier.

Classification Results: Each guideline has a dedicated classifier, and the classification performance is shown in Table 2. Overall, the accuracy of rule-based classifiers is very high (> 99.4%). Recall ranges from 75% to 100%. The variation of recall across guidelines may be attributed to a variety of factors including the number of ways to paraphrase a specific guideline, the number of available positive training cases, and the coverage of different paraphrasing patterns in the training data.

Therefore, we also used micro-averaging to combine the above values into one quantity in order to measure the classification performance across different guidelines. By micro-averaging, classifiers (for each guideline) are allowed to participate in performance evaluation equally. The micro-averaging results are listed in the last row in Table 2, with 78.3% precision, 82% recall, and 99.8% accuracy.

Quality Rating Results: The classification performance listed above attests to the effectiveness of semantics-based automated quality rating. Table 3 shows the quality scores assigned by both rule-based automated rating and human rating. Automated rating scores are pretty close to human rating scores. 45.2% of testing pages have the same score, 32.3% and19.3% of pages have one score

Table 2. Performance of sentence classification by rule-based classifiers

Rating Criteria	Human Classif -ication	Rule-based Classification (Y)	Rule-based Classification (N)	Recall	Precision	Accuracy
#1	Y	40	9	81.6%	85.1%	99.4%
	N	7	2621			
#2	Y	3	0	100.0%	42.9%	99.9%
	N	4	2670			
#3	Y	0	0	NA	NA	100.0%
	N	0	2677			
#4	Y	0	0	NA	NA	100.0%
	N	0	2677			
#5	Y	0	0	NA	NA	100.0%
	N	0	2677			
#6	Y	16	3	84.2%	84.2%	99.8%
	N	3	2655			
#7	Y	5	1	83.3%	55.6%	99.8%
	N	4	2667			
#8	Y	2	0	100.0%	100.0%	100.0%
	N	0	2675			
#9	Y	1	0	100.0%	50.0%	100.0%**
	N	1	2675			
#11	Y	6	2	75.0%	100.0%	99.9%
	N	0	2669			
#12-A	Y	9	2	81.8%	90.0%	99.9%
	N	1	2665			
#12-B	Y	10	3	76.9%	76.9%	99.8%
	N	3	2661			
#13-A	Y	4	0	100.0%	100.0%	100.0%
	N	0	2673			
#13-B	Y	2	0	100.0%	100.0%	100.0%
	N	0	2675			
#14	Y	14	4	77.8%	77.8%	99.7%
	N	4	2655			
#15	Y	2	0	100.0%	25.0%	99.8%
	N	6	2669			
#20	Y	9	3	75.0%	90.0%	99.9%
	N	1	2664			
Micro-Averaging	Y	123	27	82.0%	78.3%	99.9%
	N	34	45325			

lower or higher than human rated score. Only one page (testing page #15) has rule-based score higher by 3.

The rule-based rating results are very close to human rating results in terms of not only the number of identified criteria in each web page, but also the accuracy of identification. Given the 31 testing pages, human raters identified 92 unique guidelines. Rule-based classifiers identified 91 unique guidelines. 78

Table 3. Quality score assigned to testing web pages by rule-based auto-rating system

Testing Page ID	Quality Score	Quality Score	Quality Score
	via	via	Difference
	Human Rating	Rule-Based Rating	
1	7	7	0
2	7	6	-1
3	8	7	-1
4	6	5	-1
5	6	6	0
6	5	5	0
7	5	4	-1
8	4	5	1
9	3	4	1
10	4	3	-1
11	3	4	1
12	3	4	1
13	4	4	0
14	3	2	-1
15	2	5	3
16	2	3	1
17	2	2	0
18	2	2	0
19	2	2	0
20	3	2	-1
21	2	2	0
22	2	1	-1
23	2	1	-1
24	1	2	1
25	1	1	0
26	1	1	0
27	1	1	0
28	1	0	-1
29	0	0	0
30	0	0	0
31	0	0	0
Total	92	91	Not Applicable

identified items are the same between the two sets. 83.7% of the human identified depression treatment guidelines were also identified using the rule-based rating system. Only 16.1% of the human identified guidelines were missed; and 14.3% of computer identified guidelines were false positives.

The ultimate quality rating performance is measured using Pearson correlation between automatically rated scores and human rated scores. Given the 31 testing pages, the Pearson correlation is positive and significant, with r equal to 0.909. r^2 equals to 0.827. That means 82.7% of the variance of the rule-based quality scores is associated with the variance in the evidence-based human rated scores.

5 Discussion

Precision and recall indicate the ability of the automated approaches to correctly identify positive instances of each criterion. The higher recall, the fewer actual criteria sentences go undetected (lower false negative rate). The higher precision, the fewer non-criterion cases are mistakenly identified as a criterion (lower false positive rate). Results in Table 2 shows that the lowest recall was 75%, and the average recall was 82% across whole guideline set. This suggests that the semantics-based rule classification system can be fairly effective in identifying the presentation of depression treatment guidelines in web text.

On the other hand, Table 2 shows that precision of sentence classification varies in a wide range across different guidelines. This indicates that the classification rules defined for certain guidelines need to be enhanced with more distinguishing constraints for filtering out false positives. In addition, strongly skewed data was part of the reason for low precision. The negative over positive ratios for all criteria is averagely 302:1. Particularly, for guidelines (i.e. #2, 6 and 15) which have low precisions, the negative over positive ratio ranges from 445:1 to 1337:1. That means true positives (TP) can be easily overwhelmed by false positives (FP) even though only a very small percentage of actually negative cases are mistakenly identified as positive, hence precision can be low.

Pearson correlation is used to evaluate the quality rating scores. The strong positive correlation with human rating results suggests that the computer rated scores predicated the content quality of depression treatment web pages in a manner close to human being performance. The low precision and imperfect recall of sentence classification did not seem to have greatly affected the page rating performance. This is partly related to a fact that a single guideline is commonly paraphrased more than once in a web page. Among multiple presentation of a same guideline in a page, as long as one presentation is captured by automated classification system, the quality score is added by one without being hurt by false negatives. Similarly, suppose that the classifier label five sentences as presentations of a guideline, with four being false positives, i.e. precision = 20%. Because there is one sentence being a true positive, the impact of false positives would not been reflected in the automatically assigned quality rating score.

Although 83.7% of human identified guidelines were captured by the automated rating system, there is space for improvement to make both false positive and false negative errors lower. It is found in the case review of classification results that false positive errors occur when the semantics of a text segment is taken for the entire sentence. For example (see Figure 4), because the sentence contains “your response to certain antidepressant”, the classifier mistakenly classified the sentence as a match for guideline #1. To avoid false positives like this, the classification rules need to be supplemented with strict description logic.

Another limitation of the current implementation is that it uses individual sentences as processing unit while between-sentence analysis (e.g. co-reference) has not been utilized. For this reason, false negatives happened in a few situations in which the meaning of a guideline is expressed across multiple sentences, typically in bullet list format (see Figure 5). After sentence splitting, “exercise” and “depression” were separated into two different sentences. Hence neither of them was predicated as a presentation of guideline #20.

Guideline #1:
“Antidepressant medication is an effective treatment for major depressive disorder.”

False Positive Match:
10. The test, called the cytochrome P450, helps pinpoint genetic factors that influence your response to certain antidepressants (as well as some other medications).

Fig. 4. A false positive example

Guideline #20:
“Exercise can be effective for depression.”
False negative (missing) sentence:
Regardless of whether you have mild or major depression, the following self-care steps can help:
Get enough sleep.
Follow a healthy, nutritious diet.
Exercise regularly.

Fig. 5. A false negative example

Guideline #6: “The side effect profile varies for different antidepressants.”
Testing sentences identified as positive cases:
1. Overall, no type of antidepressant drug is more effective than any other, but the different types can have different side effects, and different drugs sometimes are more or less effective for different individuals.
2. The side effects vary depending on the type of antidepressant you take.
3. SSRIs and SNRIs are more popular than the older classes of antidepressants, such as tricyclics—named for their chemical structure—and monoamine oxidase inhibitors (MAOIs) because they tend to have fewer side effects.
4. However, because TCAs tend to have more numerous and more severe side effects, they’re often not used until you’ve tried SSRIs first without an improvement in your depression.

Fig. 6. Examples of correctly classified sentences

In spite of the limitations, the performance testing results suggest the effectiveness of the semantics-based automated quality rating approach in two aspects. First, by semantics-based classification, this study converts quality rating task to a task of identifying health care guideline presentation among web text, following a procedure similar to human rating. However, our approach is automated since computer can rely on semantics-based classification rules to distinguish positive instances (i.e. guideline presentation) from negative ones. The classification is considered semantics-based since both classification input and classification rules are generated based on semantics. Among training data, various positive training cases are semantically categorized into different groups depending on their way for paraphrasing a same treatment guideline. Features commonly existing in a same group are extracted to form a classification rule, which just corresponds to an expression pattern. Therefore, a classifier with well-trained classification rules can identify the presentation of a guideline in different expression patterns. Figure 6 shows some identified positive sentences. These examples include different ways for expressing guideline #6. Pattern a) says that side effects of antidepressants are "different"; pattern b) uses "vary" to paraphrase; pattern c) indicates variation by a discussion of "fewer/more" side effects between antidepressants. In the listed cases, the rule-based classifier successfully identified that the sentences are in concordance with the rating guideline #6.

It has to be acknowledged that training data set may not necessarily cover all expression patterns used in human communication. Thus, it is possible that a developed has incomplete classification rule set and hence can miss some positive cases in testing. Statistically speaking, however, patterns with high frequency of usage would more likely be learned from training data set than those with low frequency. Thus, the impact of missing pattern on classification recall can be controlled reasonably low by preparing the randomly sampled training data set of reasonably large size.

Second, the shallow semantic analysis and the generated semantic representation of sentences turned to be generically effective for classification tasks relative to different treatment guidelines. Through transforming text content from English natural language to semantic tag instance in our defined syntax, sentence semantics are kept and conveyed in an appropriate sufficiency for supporting classification. It is also important that the semantic tagging process in this study was independent from the treatment guidelines in that no processing was customized to deal with any specific guidelines and its unique content and concepts.

6 Conclusion

This study proposed a semantics-based approach for implementing automated quality rating on web health care pages according to evidence-based health care guidelines. Web pages with depression treatment content are used for case study. The experimental results show that the automatically generated quality scores have strong and positive correlation with human rated scores. That is, automatically generated quality scores have potential to be valid indicators of the quality of depression treatment web pages. Different from previous research, this automated approach is semantics-based, with aim to rely health care information quality rating directly on content. Through shallow semantic analysis and semantics-based classification, computer could identify the presentation of health care guidelines with reasonable accuracy (in Tables 2 and 3).

In the current implementation, the rule-based classifier utilized expression patterns manually extracted from training data to empower automated binary classification of sentences in untouched data set. Hence, human efforts for reading text and identifying the presentation of guidelines among enormous amount of web pages can be avoided. In the future we will explore the use of machine learning to enhance the automation of pattern learning process as well. In addition, we will also attempt to apply this semantics-based approach to rate the content quality of web pages in other health conditions. If the results of this study are replicable and generalizable, this automated quality rating approach could add significant value to the quality assessment practice of health care information on the web.

References

1. Barnes, C., Harvey, R., Wilde, A., et al. (2009). Review of the quality of information on bipolar disorder on the Internet. Australian and New Zealand Journal of Psychiatry, 43, 934-945.
2. Burkell, J. (2004). Health care information seals of approval: what do they signify? Information, Communication & Society, 7(4), 491-509.
3. CEBMH, (1998). CEBMH depression guideline treatment. Retrieved from http://web.archive.org/web/20040426143952/http://cebmh.warne.ox.ac.uk/cebmh /guidelines/depression/treatment.html
4. Chen, L.E., Minkes, R.K., Langer, J.C. (2000) Pediatric surgery on the Internet. Journal of Pediatric Surgery, 35, 1179-1182.
5. Crocco, A.G., Villasis-Keever, M. & Jadad, A.R. (2002). Analysis of cases of harm associated with use of health care information on the internet. Journal of the American Medical Association, 287(21), 2869-2871.
6. Eysenbach, G. Powell, J., Kuss, O. & Sa, E. (2002). Empirical studies assessing the quality of health care information for consumers on the world wide web. Journal of the American Medical Assosication, 287(20): 2691-2700.
7. Eysenbach, G. & Köhler, C. (2004). Health-related searches on the Internet. Journal of the American Medical Assosication, 291(24), 2946.
8. Frické, M. & Fallis, D. (2002). Verifiable health care information on the Internet. Journal of Education for Library and Information Science, 43(4), 246-253.
9. Frické, M., Fallis, D., Jones, M. & Luszko, G. M. (2005). Consumer health care information on the Internet about carpal tunnel syndrome: Indicators of accuracy. The American Journal of Medicine. 118, 168-174.
10. Griffiths, K.M. & Christensen, H. (2002). The quality and accessibility of Australian depression sites on the World Wide Web. The Medical Journal of Australia, 160, 97-104.
11. Griffiths, K.M. & Christensen, H. (2005) Website quality indicators for consumers. Journal of Medical Internet Research, 7(5):e55. Retrieved from http://www.jmir.org/2005/5/e55/

12. Griffiths, K.M., Tang, T.T., Hawking, D. and Christensen, H. 2005. Automated assessment of the quality of depression websites. Journal of Medical Internet Research, 7(5):e59. Retrieved from http://www.jmir.org/2005/5/e59/
13. HONcode (2012). HONcode: Principles — Quality and trustworthy health care information. Retrieved from http://www.hon.ch/HONcode/Conduct.html
14. Khazaal, Y., Chatton, A., Zullino, D. & Khan, R. (2012). HON label and DISCERN as content quality indicators of health-related websites. The Psychiatric quarterly, 83(1), 15-27.
15. Kiley, R. (2002). Does the Internet harm health? Some evidence does exist that the Internet harms health. British Medical Journal, 323(7331), 328-329.
16. Kunst, H. Groot, D., Latthe, P. Latthe, M. & Khan, K.S. (2002). Accuracy of information on apparently credible websites: Survey of five common health topics. British Medical Journal, 321(7337): 581-582.
17. Martin-Facklam, M., Kostrzewa, M., Martin, P. & Haefili, W.E. (2003). Quality of drug information on the World Wide Web and strategies to improve pages with poor information quality: An intervention study on pages about sildenafil. British Journal of Clinial Pharmacology, 57(1), 80-85.
18. Pew Internet and American Life Project. (2003). Internet health resources. Retrieved from http://www.pewinternet.org/ /media//Files/Reports/2003/ PIP_Health_ Report_July_2003.pdf
19. Pew Internet and American Life Project. (2006). Online health search 2006. Retrieved from http://www.pewinternet.org/ /media/Files/Reports/2006/ PIP_Online_ Health_2006.pdf
20. Pew Internet and American Life Project. (2011). The social life of health care information, 2011. Retrieved from http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx
21. Podichetty, V.K., Booher, J., Whitfield, M. & Biscup, R.S. (2006). Assessment of internet use and effects among healthcare professionals: a cross sectional survey. Postgraduate Medical Journal, 82:274-279
22. Silberg, W. M., Lundberg, G. D. & Musaccio, R. A. (1997) Assessing, controlling, and assuring the quality of medical information on the Internet: caveat lector et viewor–let the reader and viewer beware. Journal of the American Medical Association, 277, 1244–1245.
23. Smith, A. (2002). Evaluation of information sources. The World-Wide Web Virtual Library. Retrieved from http://www.vuw.ac.nz/staff/alastair_smith/avaln/evaln.htm
24. UMLS. (2009). UMLS Ð Metathesaurus release statistics. Retrieved from http://web.archive.org/web/20090925122534/http://www.nlm.nih.gov/research/ umls/knowledge_sources/metathesaurus/release/statistics.html
25. Wang, Y. & Liu, Z. (2007). Automatic detecting indicators for quality of health care information on the Web. International Journal of Medical Informatics,76, 575-582.