A matter of words: NLP for quality evaluation of Wikipedia medical articles

Vittoria Cozza[1], Marinella Petrocchi[1], and Angelo Spognardi[2]

[1] IIT CNR, Pisa, Italy {v.cozza, m.petrocchi}@iit.cnr.it
[2] DTU Lingby, Denmark angsp@dtu.dk

Abstract. Automatic quality evaluation of Web information is a task with many fields of applications and of great relevance, especially in critical domains, like the medical one. We move from the intuition that the quality of content of medical Web documents is affected by features related with the specific domain. First, the usage of a specific vocabulary (Domain Informativeness); then, the adoption of specific codes (like those used in the infoboxes of Wikipedia articles) and the type of document (e.g., historical and technical ones). In this paper, we propose to leverage specific domain features to improve the results of the evaluation of Wikipedia medical articles. We rely on Natural Language Processing (NLP) and dictionaries-based techniques in order to extract the biomedical concepts in a text. The results of our experiments confirm that, by considering domain-oriented features, it is possible to obtain sensible improvements with respect to existing solutions, mainly for those articles that other approaches have less correctly classified.

1 Introduction

As observed by a recent article of Nature News [10], "Wikipedia is among the most frequently visited websites in the world and one of the most popular places to tap into the world's scientific and medical information". Despite the huge amount of consultations, open issues still threaten a fully confident fruition of the popular online open encyclopedia. Among them, reliability and trustworthiness of information.

In this paper, we face the quest for quality assessment of a Wikipedia article, in an automatic way that comprehends not only reliability criteria, but also additional parameters testifying completeness of information and coherence with the content one expects from an article dealing with specific topics, plus sufficient insights for the reader to elaborate further on some argument. The notion of data quality we deal with in the paper is coherent with the one suggested by recent contributions (see, e.g., [13]), which points out like the quality of Web information is strictly connected to the scope for which one needs such information.

Our intuition is that groups of articles related to a specific topic and falling within specific scopes are intrinsically different from other groups on different topics within different scopes. We approach the article evaluation through machine learning techniques. Such techniques are not new to be employed for automatic evaluation of articles quality. As an example, the work in [18] exploits classification techniques based on structural and linguistic features of an article. Here, we enrich that model with novel features that are domain-specific. As a running scenario, we focus on the Wikipedia medical portal. Indeed, facing the problems of information quality and ensuring high and correct levels of informativeness is even more demanding when health aspects are involved. Recent statistics report that Internet users are increasingly searching the Web for health information, by consulting search engines, social networks, and specialised health portals, like that of Wikipedia. As pointed out by the 2014 Eurobarometer survey on European citizens' digital health literacy[3], around six out of ten respondents have used the Internet to search for health-related information. This means that, although the trend in digital health literacy is growing, there is also a demand for a qualified source where people can ask and find medical information which, to an extent, can provide the same level of familiarity and guarantees as those given by a doctor or a health professional.

We anticipate here that leveraging new domain-specific features is in line with this demand of articles quality. Moreover, as the outcomes of our experiments show, they effectively improve the classification results in the hard task of multi-class assessment, especially for those classes that other automatic approaches worst classify. Remarkably, our proposal is general enough to be easily extended to other domains, in addition to the medical one.

Next section describes the dataset used in our experiments. In Section 3, we introduce a domain-specific, medical model. Section 4 presents the feature extraction process, while Section 5 presents experiments and results. In Section 6, we survey related work in the area and in Section 7 we conclude the paper.

2 Dataset

We consider the dataset consisting of the entire collection of articles of the Wikipedia Medicine Portal, updated at the end of 2014. Wikipedia articles are written according to the Media Wiki markup language, a HTML-like language. Among the structural elements of one page, which differs from standard HTML pages, there are *i)* the internal links, i.e., links to other Wikipedia pages, different from links to external resources); *ii)* categories, which represent the Media Wiki categories a page belongs to: they are encoded in the part of text within the Media Wiki "categories" tag in the page source, and *iii)* informative boxes, so called "infoboxes", which summarize in a structured manner some peculiar pieces of information related the topic of the article. The category values for the articles in the medical portal span over the ones listed at `https://en.wikipedia.org/wiki/Portal:Medicine`.

Infoboxes of the medical portal feature medical content and standard coding. An infobox may contain explanatory figures and text denoting peculiar charac-

[3] `http://ec.europa.eu/public_opinion/flash/fl_404_sum_en.pdf`

teristics of the topic, such as a disease, and the value for the standard code of a disease (for example, in case of the Alzheimer's disease, the standard code is ICD9, as for the international classification[4]).

Thanks to WikiProject Medicine[5], the dataset of articles we collected from the Wikipedia Medicine Portal has been manually labeled into seven quality classes. They are ordered as *Stub, Start, C, B, A, Good Article (GA), Featured Article (FA).* The Featured and Good article classes are the highest ones: to have those labels, an article requires a community consensus and an official review by selected editors, while the other labels can be achieved with reviews from a larger, even controlled, set of editors. Actually, none of the articles in the dataset is labeled as *A*, thus, in the following, we do not consider that class, restricting the investigation to six classes.

At the date of our study, we were able to gather 24,362 rated documents. Remarkably, only a small percentage of them (1%) is labeled as *GA* and *FA*. Indeed, the distribution of the articles among the classes is highly skewed. There are very few (201) articles for the highest quality classes (FA and GA), while the vast majority (19,108) belongs to the lowest quality ones (Stub and Start). This holds not only for the medical portal. Indeed, it is common in all Wikipedia, where, on average, only one article in every thousand is a Featured one.

In Section 5, we will adopt a set of machine-learning classifiers to automatically label the articles into the quality classes. Dealing with imbalanced classes is a common situation in many real applications of classification learning: healthy patients over the population, fraudulent actions over daily genuine transactions, and so on. Without any countermeasure, common classifiers tend to correctly identify only articles belonging to the majority classes, clearly leading to severe mis-classification of the minority classes, since typical learning algorithms strive to maximize the overall prediction accuracy. To reduce the disequilibrium among the size of the classes, we have first randomly sampled the articles belonging to the most populated classes. Then, we have oversampled the data from the minority classes, following the approach in [6], the Synthetic Sampling with Data Generation. After such processing, we have 1015 articles from Start, Stub, B and C and 214 and 162 ones for GA and FA, respectively.

3 The medical domain model

We apply a multi-class classification approach to label the articles of the sampled dataset into the six WikiProject quality classes. In order to have a baseline, we have first applied the state of the art model proposed in [18] to the dataset. This model is known as the actionable model and is based on five linguistic and structural features. For page limit, we do not detail the features and how we have extracted them from the dataset. A detailed description is available in [8]. The classification results according to the baseline model are in Section 5.

[4] http://www.who.int/classifications/icd/en/

[5] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Assessment

Then, we have improved the baseline model with novel and specifically crafted features that rely on the medical domain and that capture details on the specific content of an article. As shown in Figure 1, medical model features, the biomedical entities, have been extracted from the free text only, exploiting advanced NLP techniques and using domain dictionaries.

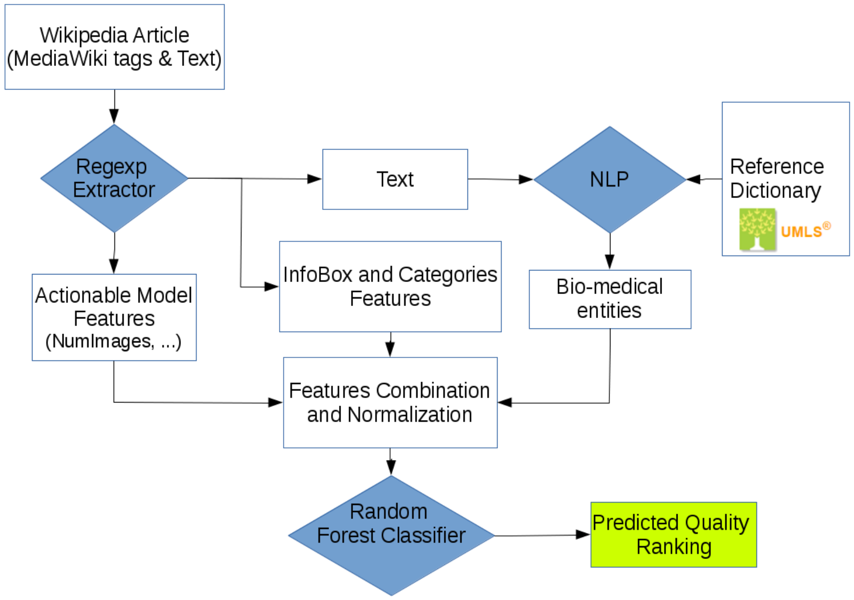

Fig. 1. Quality Assessment

In details, we newly define and extract from the dataset the following novel features:

1. *InfoBoxNormSize*: this feature represents the normalised size of an infobox that contains standard medical coding.
2. *Category*: the category a page belongs to.
3. *DomainInformativeness*: the number of bio-medical entities, which are the domain dependent terms in the article (such as the ones denoting symptoms, diseases, treatments, etc.).

The idea of considering infoboxes is not novel: for example, in [18] the authors noticed that the presence of an infobox is a characteristic featured by good articles. However, in the specific case of the Medicine Portal, the presence of an infobox does not seem strictly related to the quality class the article belongs to (according to the manual labelling). Indeed, it is recurrent that articles, spanning all classes, have an infobox, containing a schematic synthesis of the article. In particular, pages with descriptions of diseases usually have an infobox with the medical standard code of the disease (i.e., IDC-9 and IDC-10), as in Figure ??.

As done for the baseline, also the first two features of the medical model have been extracted with ad hoc Python scripts, extracting HTML structures and excerpts of the textual content within the MediaWiki tags.

For their extraction of the bio-medical entities, we consider the textual part of the article only, obtained after removing the MediaWiki tags, and we apply a NLP analysis, which is presented in Section 4.

3.1 Infobox-based feature

We have calculated the Infobox size as the base 10 log of the bytes of data contained within the mediawiki tags that wrap an infobox, and we have normalized it with respect to the ArticleLength, introduced in Section ??.

3.2 Category-based feature

We have leveraged the categories assigned to articles in Wikipedia, in particular relating to the medicine topics available at `https://en.wikipedia.org/wiki/Portal:Medicine`. We have defined 4 upper level categories of our interest:

- A *anatomy*: an article is about anatomy;
- B *biography*: an article is a biography of someone or tell the history of something;
- D *disorder*: it is about a disorder;
- F *first aid*: it reports information for first aid or emergency contacts;
- O *other*: none of the above.

We have matched the article's text within the MediaWiki categories tag with an approximate list of keywords related to our category of interest.

4 Bio-medical entities

In the literature, there are several methods available for extracting bio-medical entities from a text (i.e., from medical notes and/or articles). We refer to [12] for an overview of valuable existing techniques. In this work, we have adopted a dictionary-based approach, which exploits lexical features and domain knowledge extracted from the Unified Medical Languages System (UMLS) Metathesaurus [4]. The approach has been proposed for the Italian language in a past work [1]. Since the approach combines the usage of linguistic analysis and domain resources, we were able to conveniently adapt it for the English language, being both the linguistic pipeline and UMLS available for multiple languages (including English and Italian). The interested reader can find in [8] further details on how we have extracted the bio-medical entities.

4.1 Reference dictionary

To build a medical dictionary, we have extracted definitions of medical entities from the Unified Medical Languages System (UMLS) Metathesaurus [4]. From UMLS, we have extracted the entries belonging to the following SNOMED-CT semantic groups: *Treatment*, *Sign or Symptom*, *Disease or Syndrome*, *Body Parts, Organs, or Organ Components*, *Pathologic Function*, and *Mental or Behavioral Dysfunction*, for a total of more than one million entries, as shown in Table 1 (where the two last semantic groups have been grouped together, under *Disorder*). Furthermore, we have extracted common Drugs and Active Ingredients definitions from RxNorm[6], accessed by RxTerm[7].

semantic groups	definitions
Treatment	671,349
Sign or Symptom	43,779
Body Parts, Organs, or Organ Components	234,075
Disorder	402,298
Drugs	5,109
Active Ingredients	2,774

Table 1. Dictionary Composition

5 Experiments and results

In this section, we describe the experiments and report the results for the classification of Wikipedia medical articles into the six classes of the Wikipedia Medicine Portal. We compare the results obtained adopting four different classifiers: the actionable model in [18] and three classifiers that leverage the ad-hoc features from the medical domain discussed in the previous sections. All the experiments were realized within the Weka framework [9] and validated through 10 fold cross-validation.

For each experiment, we relied on the dataset presented in Section 2, and specifically, on that obtained after sampling the majority classes and oversampling the minority ones. The dataset serves both as training and test set for the classifiers. We have applied several classification algorithms (bagging, adaptive boosting and random forest). We report the results for the latter only.

5.1 Classifiers' features

In Table 2, we report a summary of the features for each of the considered models: the baseline model in [18] and two new models that employ the medical

[6] `https://www.nlm.nih.gov/research/umls/rxnorm/`

[7] `https://wwwcf.nlm.nih.gov/umlslicense/rxtermApp/rxTerm.cfm`

		Full	
Baseline	**Medical Domain**	**Medical Domain**	**Info Gain**
ArticleLength	ArticleLength	ArticleLength	0.939
NumHeadings	NumHeadings	NumHeadings	0.732
Completeness	Completeness	Completeness	0.724
NumRef/Length	NumRef/Length	NumRef/Length	0.621
Informativeness	Informativeness	Informativeness	0.377
	DomainInformativ.	DomainInformativ.	0.751
		InfoBoxNormSize	0.187
		Category	0.017

Table 2. Classifiers: Features and Information Gain

domain features. In the *Medical Domain* model, we add to the baseline features the Domain Informativeness, as described in Section 3 and 4. In addition, the *Full Medical Domain* model also considers the features InfoBoxNormSize and Category.

For each of the features, the table also reports the Information Gain, evaluated on the whole dataset (24,362 articles). Information Gain is a well-known metric to evaluate the dependency of one class from a single feature, see, e.g., [7].

We can observe how the Domain Informativeness feature has a considerably higher infogain value when compared with Informativeness. We anticipate here that this will lead to a more accurate classification results for the highest classes, as reported in the next section. Leading to a greater accuracy is also true for the other two new features that, despite showing lower values of infogain, are able to further improve the classification results, mainly for the articles belonging to the lowest quality classes (Stub and Start).

5.2 Classification results

Table 3 shows the results of our multi-class classification. For each of the classes, we have computed the *ROC Area* and *F-Measure* metrics [14].

At a first glance, we observe that, across all the models, the articles with the lowest classification values, for both ROC and F-Measure, are those labeled C and GA. Adding the Domain Informativeness feature produces a classification, which is slightly worse for C and FA articles, but better for the other four classes. This is particularly evident for the F-Measure of the articles of the GA class. A noticeable major improvement is obtained with the introduction of the features InfoBoxNormSize and Category in the *Medical Domain* model. The ROC Area increases for the articles of all the classes within the *Full Medical Domain*, while the F-Measure is always better than the *Baseline* and almost always better than the *Medical Domain*.

The size of an article, expressed either as the word count, analyzed in [3], or as the article length, as done here, appears a very strong feature, able to discriminate the articles belonging to the highest and lowest quality classes.

Metric	**Baseline**	**Medical Domain**	**Full Medical Domain**
ROC Area Stub	0.981	0.982	**0.983**
ROC Area Start	0.852	0.853	**0.858**
ROC Area C	0.749	0.747	**0.76**
ROC Area B	0.825	0.832	**0.836**
ROC Area GA	0.825	0.908	**0.916**
ROC Area FA	0.977	0.976	**0.978**
F-Measure Stub	0.886	**0.891**	0.89
F-Measure Start	0.587	0.582	**0.598**
F-Measure C	0.376	0.367	**0.397**
F-Measure B	0.527	0.541	**0.542**
F-Measure GA	0.245	0.338	**0.398**
F-Measure FA	0.634	0.631	**0.641**

Table 3. Classification Results (In bold, the best results)

This is testified also by the results achieved exploiting the baseline model of [18], which poorly succeeds in discriminating the articles of the intermediate quality classes, while achieving good results for Stub and FA. Here, the newly introduced features have a predominant effect on the articles of the highest classes. This could be justified by the fact that those articles contain, on average, more text and, then, NLP-based features can exploit more words belonging to a specific domain.

Then, we observe that the ROC Area and the F-Measure are not tightly coupled (namely: high values for the first metric can correspond to low values for the second one, see for example C and GA): this is due to the nature of the ROC Area, that is affected by the different sizes of the considered classes. As an example, we can observe that the baseline model has the same ROC Area value for the articles of both class B and class GA, while the F-Measure of articles of class B is 0.282 higher than that of class GA.

Finally, the results confirm that the adoption of domain-based features and, in general, of features that leverage NLP, help to distinguish between articles in the lowest classes and articles in the highest classes, as highlighted in bold in Table 3. We notice also that exploiting the full medical domain leads us to the achievement of the best results.

6 Related work

Automatic quality evaluation of Wikipedia articles has been addressed in previous works with both unsupervised and supervised learning approaches. The common idea of most of the existing work, like [16, 3, 20, 19, 18], is to identify a feature set, having as a starting point the Wikipedia project guidelines, to be exploited with the objective in mind to automatically label the articles.

Recent studies specifically address the quality of medical information. In [2], the authors debate if Wikipedia is a reliable learning resource for medical students, evaluating articles on respiratory topics and cardiovascular diseases. In [11] the authors provide novel solutions for measure the quality of medical information in Wikipedia, by adopting an unsupervised approach based on the Analytic Hierarchy Process, a multi-criteria decision making technique [15]. The work in [5] aims to provide the web surfers a numerical indication of Quality of Medical Web Sites. A similar measurement is considered in [17], where the authors present an empirical analysis that suggests the need to define genre-specific templates for quality evaluation and to develop models for an automatic genre-based classification of health information Web pages. In addition, the study shows that consumers may lack the motivation or literacy skills to evaluate the information quality of health Web pages. Clearly, this further highlights the cruciality to develop accessible automatic information quality evaluation tools and ontologies. Our work moves towards the goal, by specifically considering domain-relevant features and featuring an automatic classification task spanning over more than two classes.

7 Conclusions

In this work, we aimed to provide a fine grained classification mechanism for all the quality classes of the articles of the Wikipedia Medical Portal. An important and novel aspect of our classifier, with respect to previous works, is the leveraging of features extracted from the specific, medical domain, with the help of Natural Language Processing techniques. As the results of our experiments confirm, considering specific domain-based features, like Domain Informativeness and Category, can eventually help and improve the automatic classification results. We are planning to extend the work to include other domains, in order to further validate our approach.

References

1. G. Attardi, V. Cozza, and D. Sartiano. Adapting linguistic tools for the analysis of italian medical records. *Italian Conference on Computational Linguistics CLiC-it 2014*, 2014.
2. S. A. Azer. Is Wikipedia a reliable learning resource for medical students? Evaluating respiratory topics. *Advances in Physiology Education*, 39(1):5–14, 2015.
3. J. E. Blumenstock. Size matters: Word count as a measure of quality on Wikipedia. In *17th World Wide Web*, pages 1095–1096. ACM, 2008.
4. O. Bodenreider and A. T. McCray. Exploring semantic groups through visual approaches. *Journal of biomedical informatics*, 36(6):414–432, 2003.
5. F. Cabitza. An information reliability index as a simple consumer-oriented indication of quality of medical web sites. In *Quality Issues in the Management of Web Information*, volume 50, pages 159–177. Springer, 2013.
6. N. V. Chawla et al. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

7. T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.
8. V. Cozza, M. Petrocchi, and A. Spognardi. A matter of words: NLP for quality evaluation of Wikipedia medical articles. *CoRR*, abs/1603.01987, 2016.
9. M. Hall et al. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.
10. R. Hodson. Wikipedians reach out to academics. *Nature News*, Sept. 2015.
11. E. Marzini et al. Improved automatic maturity assessment of Wikipedia medical articles. In *ODBASE*, pages 612–622. Springer, 2014.
12. P. Nakov and T. Zesch, editors. *Semantic Evaluation (SemEval)*. Association for Computational Linguistics, August 2014.
13. G. Pasi et al. An introduction to quality issues in the management of web information. In *Quality Issues in the Management of Web Information*, volume 50 of *Intelligent Systems Reference Library*, pages 1–3. Springer, 2013.
14. D. M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technologies*, 2(1):37–63, 2011.
15. T. L. Saaty. How to make a decision: The Analytic Hierarchy Process. *European Journal of Operational Research*, 48(1), 1990.
16. B. Stvilia et al. Assessing information quality of a community-based encyclopedia. In *Information Quality*, pages 442–454, Cambridge, MA, 2005. MIT.
17. B. Stvilia et al. A model for online consumer health information quality. *American Society for Information Science and Technology*, 60(9):1781–1791, 2009.
18. M. Warncke-Wang et al. Tell me more: An actionable quality model for Wikipedia. In *9th Symposium on Open Collaboration*, pages 8:1–8:10. ACM, 2013.
19. K. Wecel and W. Lewoniewski. Modelling the quality of attributes in Wikipedia infoboxes. In *Business Information Systems Workshops*, volume 228 of *Business Information Processing*, pages 308–320. Springer International Publishing, 2015.
20. K. Wu et al. Mining the factors affecting the quality of wikipedia articles. *Information Science and Management Engineering*, 1:343–346, Aug 2010.