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

Vittoria Cozza, Marinella Petrocchi and Angelo Spognardi
Wikipedia and WikiProject medicine

- Wikipedia:
  - the most popular online encyclopedia
  - tapping into the world’s scientific and medical info
  - one of the most visited websites

- Around six out of ten respondents have used the Internet to search for health-related information [Eurobarometer, updated late 2014]
- Wikipedia includes several medical articles under the WikiProject medicine portal
- Wikipedia suffers from trustworthiness issues
- Data quality and appropriate levels of informativeness are even more demanding when health aspects are involved
Wikipedia bots

- Bots act as real users and take care of article creation and editing
- Examples

  User:ClueBot NG – reverts vandalism
  User:CorenSearchBot – checks for copyright violations on new pages
  User:Lowercase sigmabot III – archives talk pages

- For a full list: https://en.wikipedia.org/wiki/Wikipedia:Bots
Towards Wikipedia Smart Bots

- Automatic quality assessment
- Vandalism detection
- Opinion spamming and spammer detection
Guidelines for Quality Assessment

- A number of English Wikipedia articles have been manually evaluated along with a quality label in Wikimedia project
- Guidelines consider linguistic, structural, historical, reputational criteria
- Stub, Start, C, B, A, Good Article (GA), Featured Article (FA)
- GA / FA require a community consensus and a social review by selected editors
Automatic Quality Assessment

- Stvilia et al. (2009):
  - linguistic (i.e., Flesch reading-ease score, structural, historical and reputational)
  - clustering and classification to detect FA (90% correctly identified)

- Blumenstock (2008): word count is the most discriminative in identify FA vs others.

Stvilia (2009). A model for online consumer health information quality. JASIST
WWW 2008
Baseline: Actionable model

- **Actionable Model** [Wang 2013], with features related to the content of articles
- The model can also **directly suggest strategies for improving** a given article quality:
  - Completeness = 0.4*NumBrokenWikilinks + 0.4*NumWikilinks
  - Informativeness = 0.6*InfoNoise + 0.3*NumImages
  - NumHeadings
  - ArticleLength
  - NumReferences/ArticleLength
- **Classifiers**: Bagging, ADA Boosting, Random Forest

Dataset

- Dec. 2014: 24,362 rated documents
- very few (201) articles for FA and GA
- vast majority (19,108) are in the lowest quality classes (Stub and Start)
- we sampled the majority classes
- and oversampled the minority classes
- labeled dataset -> supervised approach
Medical Domain model: Quality Assessment process

Wikipedia Article (MediaWiki tags & Text)

Regexp Extractor

Actionable Model Features (NumImages, ...)

InfoBox and Categories Features

Features Combination and Normalization

Random Forest Classifier

Text

NLP

Bio-medical entities

Reference Dictionary

UMLS®

Predicted Quality Ranking
InfoBox Feature

- Correlation between the quality of an InfoBox and the article quality itself: it’s a characteristic featured by GA[1]
- InfoBoxes are strongly correlated to entity types
- InfoboxBoxNormSize is the log10 of the bytes of data contained within the MediaWiki tags that wrap an infobox, normalized to the article length

Categories Feature

- We extracted the article category of interest as:
  - A, when an article is about anatomy;
  - B, when an article is a biography or an event relevant for medicine;
  - D, if it is about a disorder;
  - F, when it is about first aid or emergency contacts;
  - O otherwise

Extraction by matching the text within the categories tags with a list of keywords in our categories of interest
Domain Informativeness

- Number of bio-medical entities (e.g., symptoms, diseases, treatments, etc.)
- Bio-medical entities extraction:
  - application of NLP analysis to the textual part of the article
  - Adoption of a dictionary-based approach
Bio-medical Entity Extraction 1/3

• Dictionary based approach:
  • A large unlabeled text
  • Preliminary linguistic analysis (sentence splitting, tokenization, lemmatization, Part Of Speech Tagging):
    • UniPi Tanl Linguistic pipeline(*)
  • A reference dictionary

(*http://tanl.di.unipi.it/en/)
We created an English medical Thesaurus for medical documents, by extracting definitions from UMLS metathesaurus:

- Definition included in SNOMED CT (core terminology for EHR)
- Active Ingredients and Drugs from RxNorm
- more than one million entries:

<table>
<thead>
<tr>
<th>semantic groups</th>
<th>definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>671,349</td>
</tr>
<tr>
<td>Sign or Symptom</td>
<td>43,779</td>
</tr>
<tr>
<td>Body Parts, Organs, or Organ Components</td>
<td>234,075</td>
</tr>
<tr>
<td>Disorder</td>
<td>402,298</td>
</tr>
<tr>
<td>Drugs</td>
<td>5,109</td>
</tr>
<tr>
<td>Active Ingredients</td>
<td>2,774</td>
</tr>
</tbody>
</table>
Bio-medical Entity Extraction 3/3

• Identification of n-grams, with 1<=n<=10, in a sentence and matching them with definitions in the reference dictionary
  • Exact Match
  • Approximate match:
    • considering the lemmas
    • not considering punctuation, prepositions and articles

Example
«Other risk factors include a history of head injuries, depression, or hypertension»

Head injuries matches with head injury in the dictionary, even if word number differs
Experiments & Results

- 3 models
- Full Medical Domain with ALL NEW features
- Medical Domain with DomainInformativeness
- State of art Actionable Model
Experiments & Results

• Best results obtained with
• Random Forest Classifier trained with the selected data, wrt 6 quality classes
• 10 cross folder validation

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

- A fine grained classification for all the quality stages of the articles in Wikimedia Medicine Portal.
- **NOVELTY**: NLP techniques for quality assessment.
- Approach adaptable to other languages and other domains
- **Full Medical Domain** outperforms the baseline for high quality classes, especially GA
## Who’s Who

<table>
<thead>
<tr>
<th>Author</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vittoria Cozza</td>
<td>Polytechnic University of Bari, Italy</td>
</tr>
<tr>
<td>Marinella Petrocchi</td>
<td>IIT CNR, Pisa, Italy</td>
</tr>
<tr>
<td>Angelo Spognardi</td>
<td>DTU Compute Lingby, Denmark</td>
</tr>
</tbody>
</table>