6-9 June 2016, USI Lugano, Switzerland

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

Vittoria Cozza, Marinella Petrocchi and Angelo Spognardi

Politecnico di Bari

Istituto di Informatica e Telematica del CNR

Technical University of Denmark

Wikipedia and WikiProject medicine

- Wikipedia:

the most popular online encyclopedia

tapping into the world's scientific anf medical info

one of the most visited webistes

[Alexa.com]

- 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

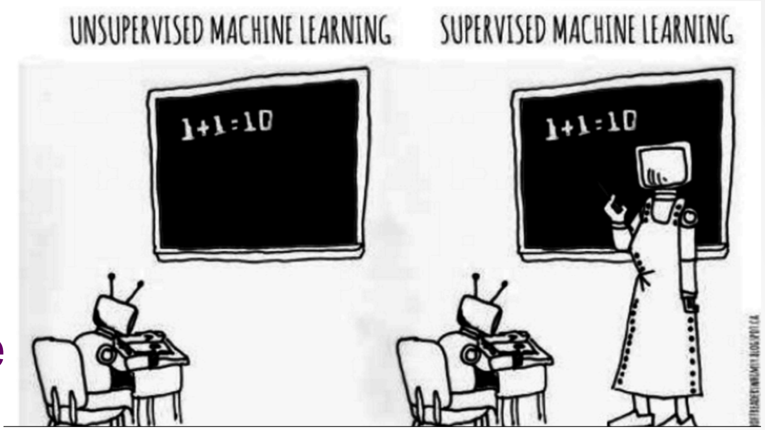
Towards Wikipedia Smart Bots

- ☑ Automatic quality assessment
- ❑ Vandalism detection
- ❑ Opinion spamming e opinion 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
Blumenstock (2008). Size matters: Word count as a measure of quality on Wikipedia. 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

Wang et al.: Tell me more: an actionable quality model for Wikipedia, Wikisym (2013)

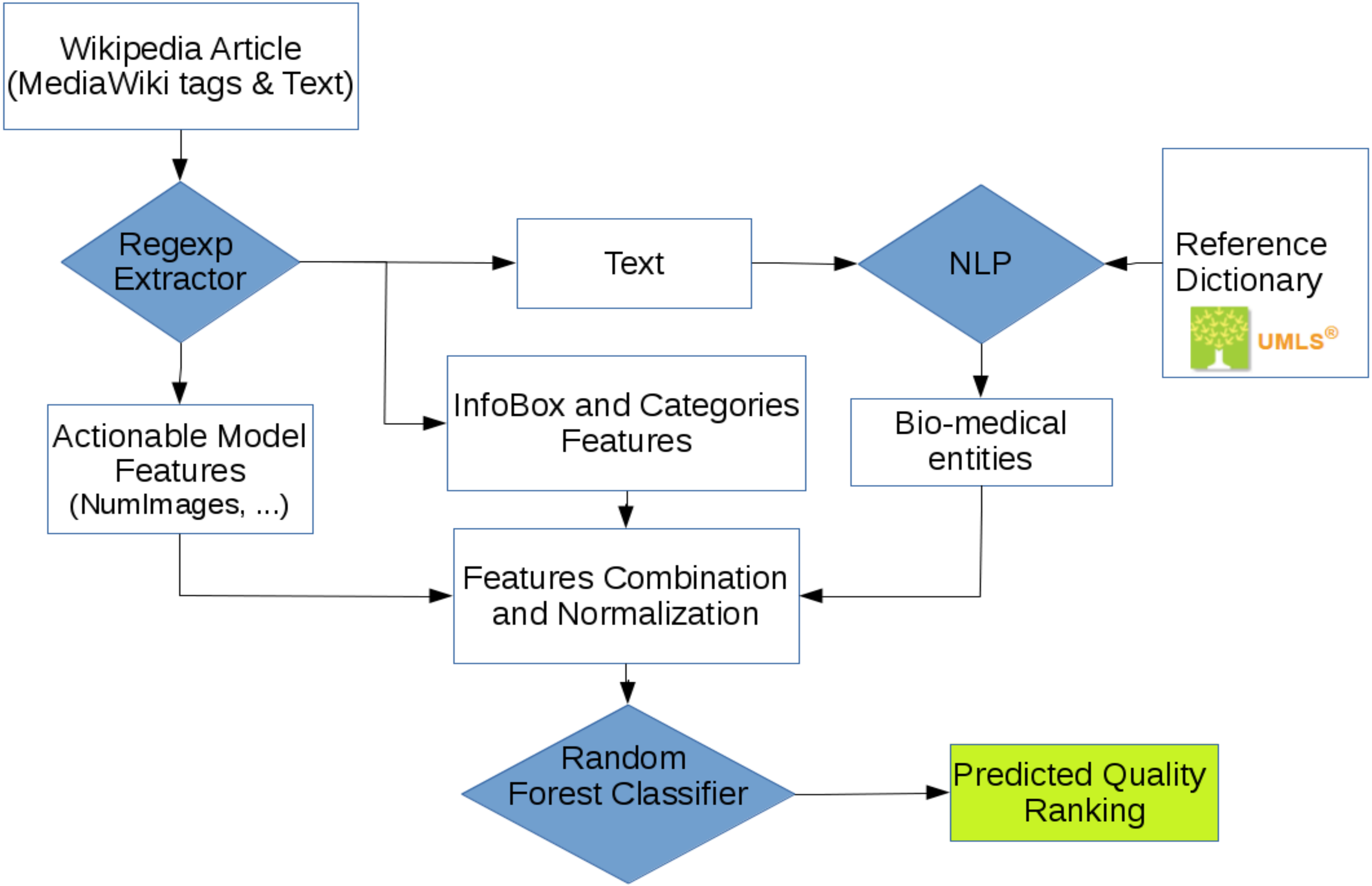
Dataset

class	original dataset	with majority classes sampling	with minority classes oversampling
Stub	9,267	1,015	1,015
Start	9,841	1,015	1,015
C	3,149	1,015	1,015
B	1,894	1,015	1,015
GA	153	153	214
FA	58	58	162
total	24,362	4,271	4,436

Table 1. Dataset

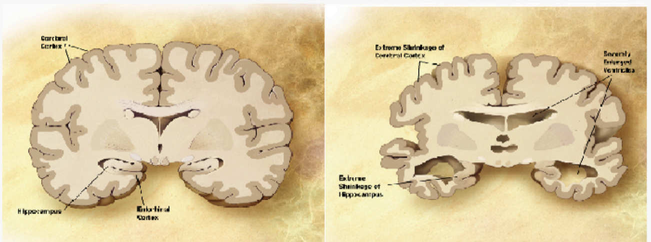
- Dec. 2014: 24,362 rated documents
- very few (201) articles for FA and GA
- vast majority (19,108) are in the lowest quality classes (Stub and Start)
- we sampled the majority classes
- and oversampled the minority classes
- labeled dataset -> supervised approach

Medical Domain model: Quality Assessment process



InfoBox Feature

Alzheimer's disease



Comparison of a normal aged brain (left) and the brain of a person with Alzheimer's (right). Characteristics that separate the two are pointed out.

Classification and external resources

Specialty	Neurology
ICD-10	G30, F00
ICD-9-CM	331.0, 290.1
OMIM	104300
DiseasesDB	490
MedlinePlus	000760
eMedicine	neuro/13
Patient UK	Alzheimer's disease
MeSH	D000544
GeneReviews	NBK1161

- 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

[1] Krzysztof Węcel , Włodzimierz Lewoniewski. «Modelling the Quality of Attributes in Wikipedia Infoboxes» Business Information Systems Workshops

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

Categories: Alzheimer's disease | Ailments of unknown etiology | Unsolved problems in neuroscience | Learning disabilities | Psychiatric diagnosis | Dementia | Abnormal psychology | Cognitive disorders | Aphasias | Herpes simplex virus-associated diseases

D

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

category	list of keywords
A	anatom*, embryolog*, organ, tissue
B	born, death, birth
D	disorder, disease, pathology
F	first aid

Table 2. Categories

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

Form	Lemma	POS
Other	other	JJ
risk	risk	NN
factors	factor	NNS
include	include	VBP
a	a	DT
history	history	NN
of	of	IN
head	head	NN
injuries	injury	NNS
,	,	,
depression	depression	NN
,	,	,
or	or	CC
hypertension	hypertension	NN
.	.	.

[1] Attardi, Cozza, Sartiano. «Adapting Linguistic Tools for the Analysis of Italian Medical Records» CLiC-it 2014
(*)http://tanl.di.unipi.it/en/

Bio-medical Entity Extraction 2/3

- 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:

Unified Medical Language System

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

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 puntuaction, 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

Baseline	Medical Domain	Full 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

Experiments & Results

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

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**

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

Author	Affiliation
Vittoria Cozza	Polytechnic University of Bari, Italy
Marinella Petrocchi	IIT CNR, Pisa, Italy
Angelo Spognardi	DTU Compute Lingby, Denmark