Automatically Identifying Health- and Clinical-Related Content in Wikipedia

Feifan Liu[1], Sohei Moosavinasab[2], Shashank Agarwal[3], Andrew S. Bennett[4,5], **Hong Yu[6,7]**

1. NLP R&D, Nuance Communications Inc
2. University of Wisconsin Milwaukee
3. DataXu Inc
4. Medical College of Wisconsin; 5. Clement J Zablocki VA medical center
6. University of Massachusetts Medical School; 7. VA Central Western MA

Outline

- Motivation
- Related work
- Task definition and Methods
  - Unsupervised method
  - Supervised method
- Experimental Results
- Conclusion and future work

Motivation

- AskHERMES is a clinical question answering system
  - It typically indexes trusted knowledge resources
- Wikipedia has shown to be an important medical information resource
  - 70% of junior physicians use Wiki in a given week, about 50% to 70% of practicing physicians use it as an information source
  - 35% pharmacists uses Wiki for medical information
- Can we integrate Wikipedia into the AskHERMES system?

Motivation(Cont.)

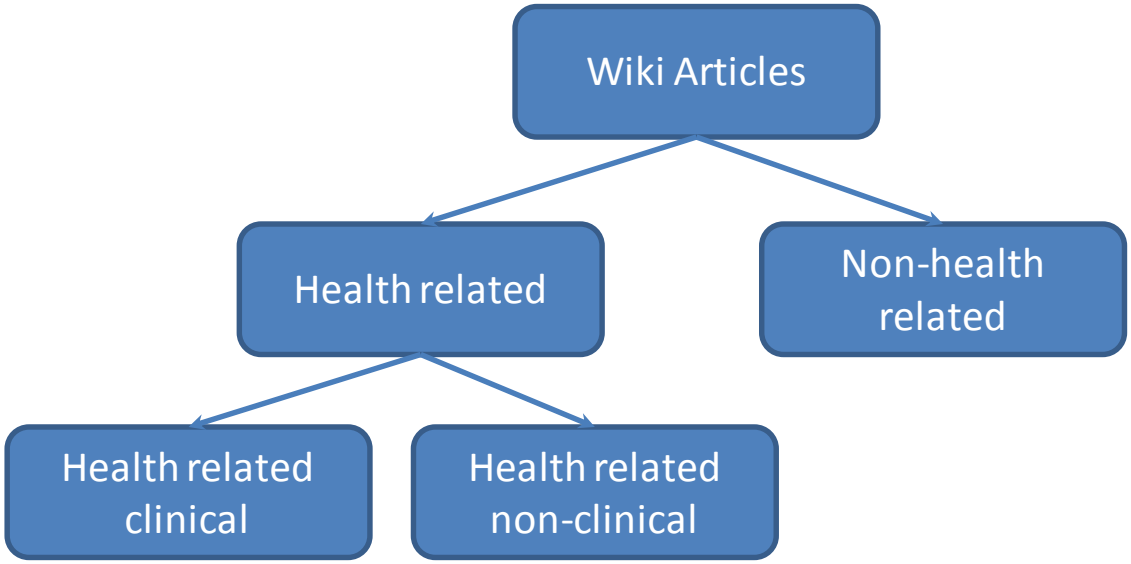
- Challenges
  - Wikipedia contains a **full-spectrum** of world knowledge
    - Ambiguities across health domain and other domains, making IR module in QA more susceptible to get spurious relevant documents
  - Categories assigned to Wiki articles, by themselves, don't suffice to reliably predict domain semantics
    - Noisy hierarchy
- Our work
  - Pilot effort to identify health related content from Wikipedia
  - Exploitation of category tags in an alternative way
  - First step towards optimal integration between clinical QA and Wikipedia

Related Work

- To the best of our knowledge, no studies made an effort to integrating Wikipedia into clinical QA
- More work shows Wikipedia becomes more and more popular in the clinical domain
  - [Friedlin et al 2010] Effective knowledge base for medical informatics
  - [Rajagopalan et al 2011] similar accuracy and depth as professionally edited database
  - [Reavley et al 2012] high quality information on depression and schizophrenia from Wikipedia
- Automatic assignment of Wikipedia category label [Szymanski 2010]
  - Too fine-grained label
- Use Wiki category hierarchy for different applications
- We investigate how category hierarchy can help identify health and clinically related content from Wikipedia

Task Definition

- Classification Task

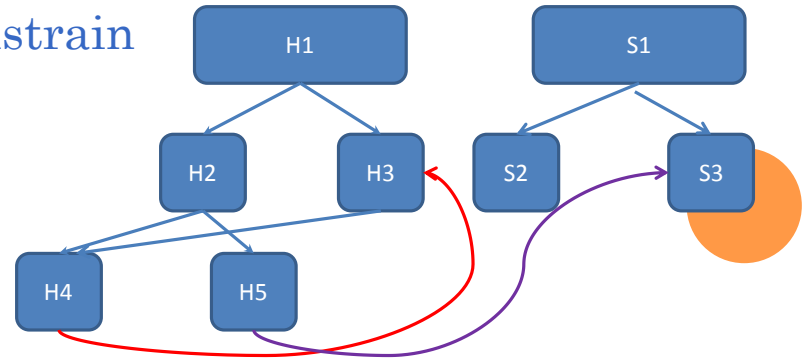


Methods

- Unsupervised
- Supervised

Unsupervised Method

- Wiki Category Hierarchy
  - 22 main topic categories in the top level hierarchy
    - "Health" is one of them
  - Assumption
    - Category labels under Health convey "health related" semantics
    - the more health related categories that are assigned to an article, the more likely this article is health-related
- Challenge of Extracting the Health Hierarchy
  - Not a tree structure and thus contains cycles → redundancy even failure
  - Some categories have multiple parent categories → noise
- Solution
  - Breath-first traverse with level constrain



Unsupervised Method(Cont.)

- Health Categories Percentage(hcp)
  - Percentage of health related categories assigned to an article
  -

$$hcp = \frac{\#\ of\ \ health\ categories}{total\ \#\ of\ \ categories\ of\ \ an\ article}$$

- Decision Function

$$f(w,t) = \begin{cases} 1(health) & if\ \ hcp > t \\ 0(non-health) & otherwise \end{cases}$$

- Issues
  - Threshold t has to be empirically selected
  - More vulnerable to noisy categories

Supervised Approach

- Four classification scenarios
  - Binary classification: health vs. non-health
  - Binary classification: health clinical vs. health non-clinical
  - Multiclass classification: health clinical vs. health non-clinical vs. non-health
  - Pipeline system: connect the first two together

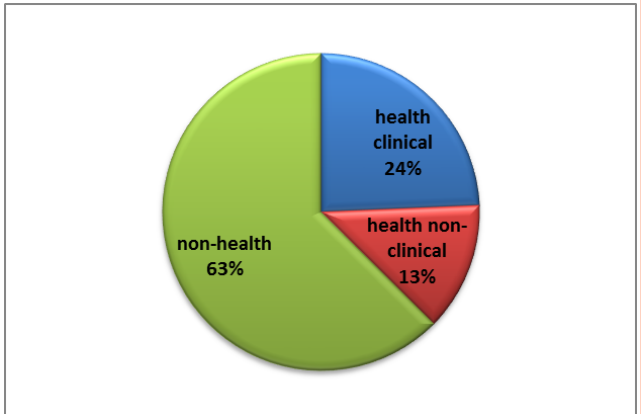
- Features explored
  - Bag of words
    - Unigram in different sections with text normalization
  - Wikipedia category features
    - Names of assigned health categories
    - Parent category names in the health category hierarchy
    - Health category percentage
    - Statistics on depth levels of assigned health categories
      - Max, min, avg, std

Supervised Approach(Cont.)

- Learning model
  - Naïve Bayes Multinomial
  - Used Weka machine learning toolkit
- Feature selection
  - Mutual information with different thresholds

Gold Standard Data

- Data source
  - Wikipedia database dump file (20120601)
- Annotation
  - 9 annotators were recruited
  - Each annotator annotated 120 articles with 40 articles overlapping with two other annotators(20 for each)
    - 720 were annotated once, 180 articles were annotated twice
- Final
  - Exclude 14 (out of 720) with the certainty level “not sure”
  - Exclude 31 (out of 180) where two annotators has the same certainty level but differ in annotations
  - Gold standard have 855 articles



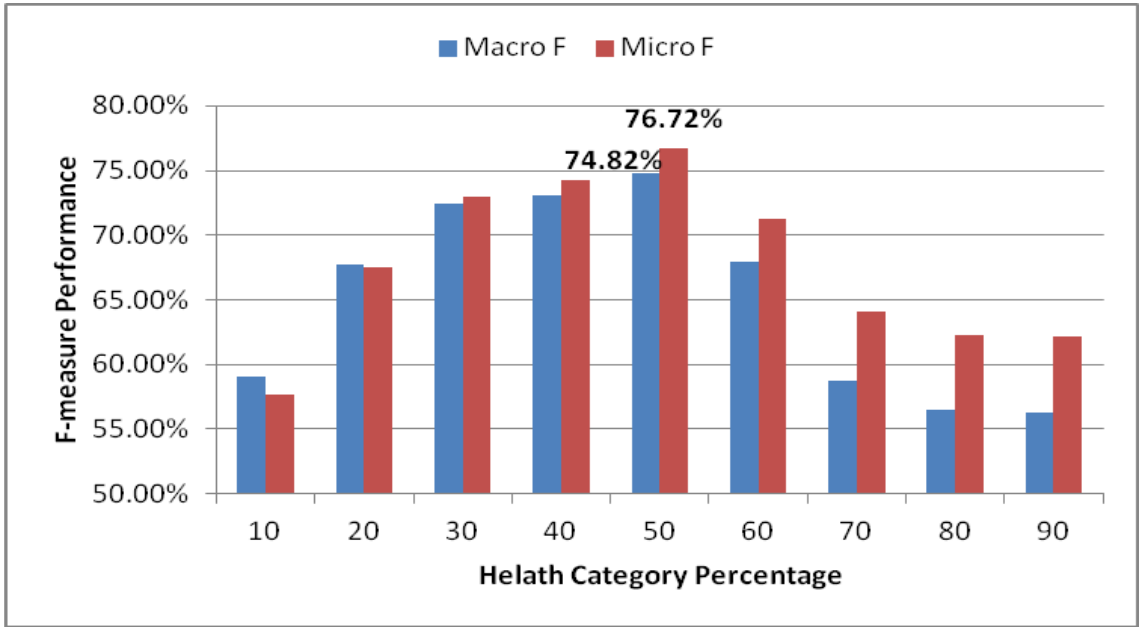
Inter-agreement on annotation

KAPPA COEFFICIENT

Categories evaluated	Article coverage	
	All articles	Only sure articles
All three categories	0.595	0.649
Health/Non-health	0.664	0.721
Health related clinical/ Health related non-clinical	0.535	0.615

Unsupervised Result

- Results based on different thresholds



Supervised Result(1)

Features Used	W/O FS		With FS	
	Macro F1 (%)	Micro F1 (%)	Macro F1 (%)	Micro F1 (%)
BOW	85.68 ± 3.84	86.68 ± 3.56	86.04 ± 4.52	86.91 ± 4.23
BOW+ hcp	81.20 ± 2.68	81.75 ± 2.83	86.56 ± 2.98	87.32 ± 2.80
BOW + cat	85.38 ± 3.71	86.42 ± 3.43	86.17 ± 4.55	87.03 ± 4.26
BOW + parent	87.40 ± 4.55	88.24 ± 4.23	87.26 ± 4.70	88.12 ± 4.37
BOW + cat+parent	**87.91 ± 4.53**	**88.72 ± 4.23**	**87.91 ± 4.53(all)**	**88.72 ± 4.23(all)**
BOW + level	85.05 ± 4.18	86.14 ± 3.85	86.31 ± 4.64	87.14 ± 4.33
Bag of words + all_cat	85.94 ± 2.97	86.56 ± 2.96	87.89 ± 2.70	88.64 ± 2.52

Supervised Result(2)

- Health vs. Non-health

Features Used	W/O FS		With FS	
	Macro F1 (%)	Micro F1 (%)	Macro F1 (%)	Micro F1 (%)
BOW	84.21 ± 7.29	85.99 ± 6.45	85.36 ± 7.69	87.14 ± 6.45
BOW + cat+parent	**88.06 ± 4.91**	**89.36 ± 4.40**	**88.06 ± 4.91**	**89.36 ± 4.40**

- Best performance in three way classification: Non-Health, Health-Clinical and Health-non-Clinical

	Macro F1 (%)	Micro F1 (%)
Multiclass	**78.38**	**85.37**
Pipeline	75.71	83.91

Error Analysis

- "Patrick Mullie"
  - Annotated as "Non"health" and the system get it right as "Health"
  - Annotated as "Non health" and the system get it right as "Health"
- "List of disorders of foot and ankle"
  - Annotated as "health non-clinical" and the system classified it correctly as "health clinical"
  - "List of disorders of foot and ankle"
  - Annotated as "health non-clinical", and the system classified it correctly as "health clinical"
- "Total petroleum hydrocarbon": the system detected it as "Health clinical" in correctly, as it contains chemical which is common in drug names
  - Bag of words lacks deep semantics
  - "total petroleum hydrocarbon": the system detected it as "Health clinical" in correctly, as it contains chemical which is common in drug names
- "Damping off" is a plant disease, but the system got it as "health related clinical", failed to infer it is not a

Conclusion and Future Work

Automatic learning to identify health related content

- Explored both textual features and features from wiki category hierarchy
- Promising results were obtained
- Future work

- More features will be explored
  - Syntactic parsing
    - hyperlinks
    - hyperlinks
  - Incorporate the system into clinical QA systems
- Incorporate the system into clinical QA systems

Acknowledgement

Hong Yu

Sciences of the National Institute of Health
National Center for Advancing Translational
Massachusetts Medical School to Hong Yu
This research was supported by UL1TR000161

- National Center for Advancing Translational Sciences of the National Institute of Health under award number UL1TR000161.

THANK YOU!