Discovering Hidden Knowledge from Scientific Literature: Challenges and Some Solutions

Mark Stevenson

Natural Language Processing Group
University of Sheffield, UK

http://staffwww.dcs.shef.ac.uk/people/M.Stevenson/

@drmarkstevenson

Outline

- Hidden knowledge in Scientific Literature
- Challenges in Identifying Hidden Knowledge
  1. Volume
  2. Ambiguity
  3. Inconsistency

Hidden Knowledge

- Hidden knowledge occurs when a connection can be inferred by combining information in multiple documents but that connection has not been noticed.

- **Literature Based Discovery (LBD)** has been used to discover hidden knowledge

Hidden Knowledge Examples

- Raynaud's Syndrome can be treated using Fish Oil (Swanson, 1986)
- Magnesium deficiency can cause migraine headaches (Swanson, 1988)
- Indomethacin can treat Alzheimer's Disease (Smallheiser and Swanson, 1996)

Work on LBD

- "Language Processing for Literature Based Discovery in Medicine" EPSRC (2012-5)
  - With Sheffield's Medical School (Neuroscience and Oncology)

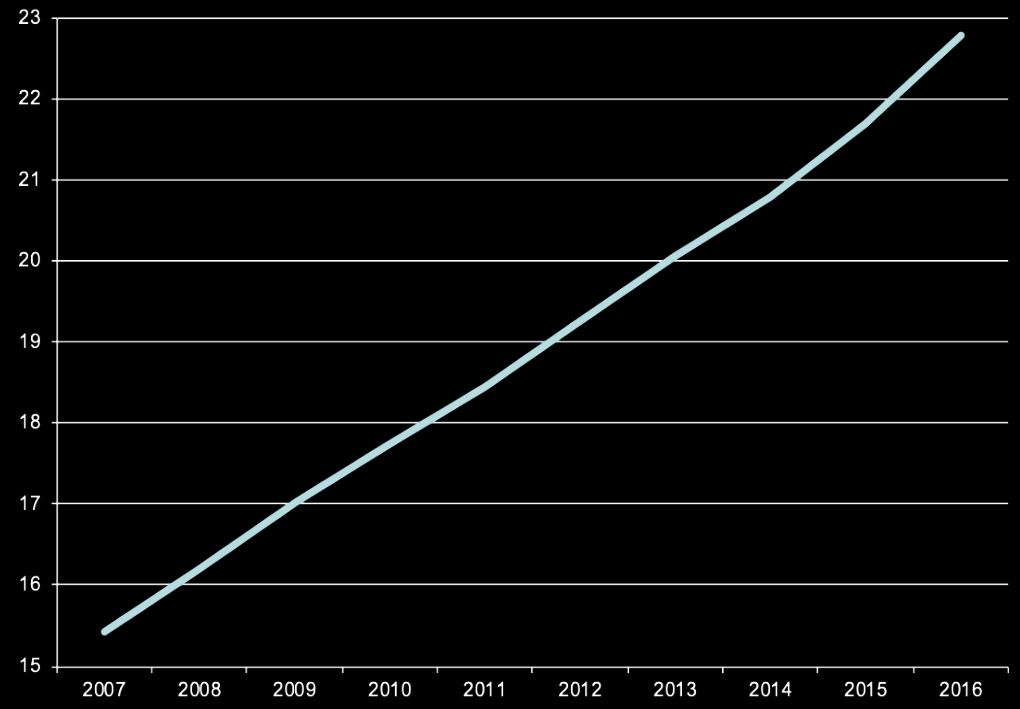
- "HypoGen: Hypothesis Generation and Visualisation from Large Corpora", Defence Science and Technology Laboratories (2016)

Some Challenges

1. Volume
2. Linguistic Ambiguity
3. Inconsistency

Challenge 1: Volume

- The literature on biomedicine and the life sciences is vast and growing rapidly



Volume and LBD

- Volume of hidden knowledge candidates that can be generated grows polynomially on the size of the input corpus.
  - Straightforward to generate more candidates than can feasibly be explored.
- Choice of how connections are defined has important effect on the number of candidates generated.

Experiment

- Six approaches to identifying connections
- Three based on co-occurrence:
  1. **c-doc**: terms co-occur in an abstract
  2. **c-sent**: terms co-occur in a sentence
  3. **c-title**: terms co-occur in a title
- Three based on linguistic analysis:
  1. **SemRep**
  2. **ReVerb**
  3. **Stanford**

Simulating LDB

- Evaluation using "time-slicing"

	Hidden Knowledge	Correct	F-measure
c-doc	14,601,340,987	762,474	1.04e-04
c-sent	5,697,603,946	1,104,869	3.88e-04
c-title	786,977,001	1,392,441	3.53e-03
SemRep	197,590,213	1,268,934	1.27e-02
ReVerb	162,065,341	1,068,498	2.28e-02
Stanford	74,442,449	885,203	**2.32e-02**

Replication of Existing Discoveries

	SemRep	ReVerb	Stanford
RD – fish oil	4	0	1
Somatomedin C – Arg	130	22	27
Migraine – Mg	47	3	13
Mg deficiency – ND	43	5	0
AD – estrogen	331	64	76
AD – INN	234	47	49
Schizophrenia – Ca2+iPLA2	13	0	0

- No results for co-occurrence approaches given volume of links. (c-doc & c-sent guaranteed to replicate discoveries)

Challenge 2: Ambiguity

- Ambiguity can make LBD difficult
- Mg can mean Magnesim or Milligram
  - Caused problems in replicating magnesium/ migraine connection (Weeber, 2001)

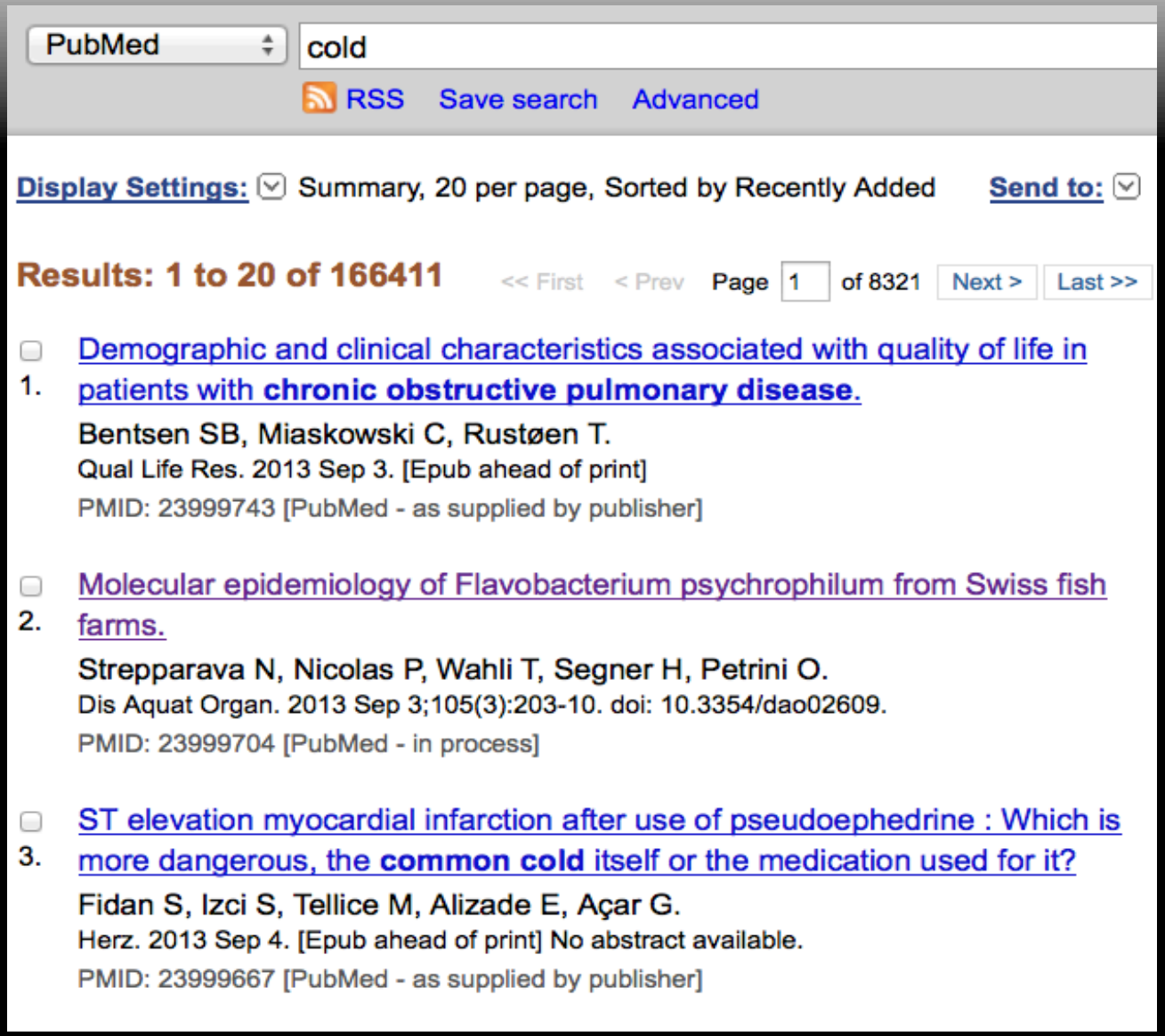
Ambiguity in Biomedical Documents

- Generally believed that ambiguities do not occur with domains

- One Sense per Discourse (Gale, Church and Yarowsky, 1992)
  - "there is a very strong tendency (98%) for multiple uses of a word to share the same sense in a well-written discourse"

Ambiguity in Medline: cold

- Cold temperature
  - "generation of suppressor macrophages during acute cold stress" (PMID 9338419)
- Common cold
  - "personal histories of hypertension and thyroid disease, and susceptibility to colds" (PMID 9251855)
- COLD (Chronic Obstructive Lung Disease)
  - "fifty-six smoking patients with COLD" (PMID 9411973)



Chronic Obstructive Lung Disease

temperature

common cold

Other Ambiguities

- Culture
  - Laboratory culture: "peripheral blood mononuclear cell culture" (PMID 9363902)
  - Anthropological culture: "A cross-cultural breakdown" (PMID 9272194)
- Transport
  - Biological transport: "glutamate is transported into Muller cells" (PMID 9695799)
  - Patient transport: "complications associated with transporting critically ill patients" (PMID 9674486)

Experiment

	WSD performance (F-measure)	# pairs	LBD performance (Scaled F-measure)
WSD-1	46.8%	4,554,466,783	1.00
WSD-2	32.4%	175,748,768	0.38
WSD-3	29.5%	162,065,341	0.35
Random	29.3%	133,004,828	0.29

WSD for Biomedical Documents

- Explored a range of approaches:
  - Unsupervised graph-based approach
    - Convert UMLS into a graph and use the Personalised PageRank algorithm to disambiguate
  - Supervised approaches
    - Automatic generation of training examples

DALE system

- Created using labeled examples generated for 103,929 CUIs (2010AA UMLS version)

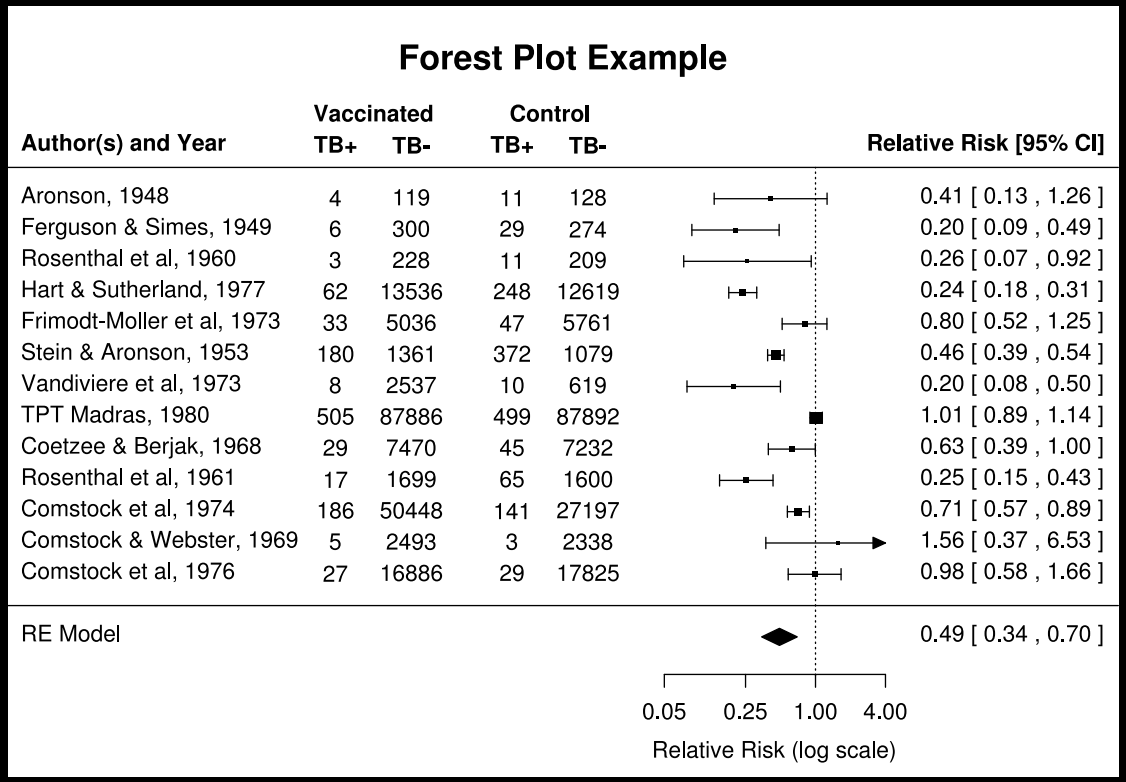
http://kta.rcweb.dcs.shef.ac.uk/dale/

Challenge 3: Inconsistency

- Does aspirin cause bleeding in patients undergoing coronary bypass surgery?
  - "Mean mediastinal blood loss was significantly greater in the aspirin group (919 +/- 164 ml., S.E.) than in the control group (437 +/- 61 ml., p less than 0.001)." (PMID: 309032)
  - "Patients taking 85-325 mgm of aspirin with a normal bleeding time undergoing elective CABG did not have increased RBC loss or increased transfusion requirements." (PMID: 2010437)

Systematic Reviews

- Summarise evidence answering research question
- Useful for identifying contradictory claims



Contradiction Corpus

- Contains 259 studies derived from 24 systematic reviews
- Annotated with claim, polarity and claim type (kappa 0.67 – 0.94)

```
<REVIEW REVIEW_PMID="24212980" REVIEW_TITLE="Safety and Efficacy Outcomes of
Preoperative Aspirin in Patients Undergoing Coronary Artery Bypass Grafting: A Systematic Review and
Meta-Analysis.">
 – <CLAIM ASSERTION="YES" PMID="15888837" QUESTION="In patients undergoing coronary bypass
   surgery, does Aspirin usage, compared to no aspirin, cause bleeding" TYPE="CAUS">
      The administration of aspirin until the operation may improve oxygenation with only a slight increase in
      bleeding.
   </CLAIM>
 – <CLAIM ASSERTION="NO" PMID="21509719" QUESTION="In patients undergoing coronary bypass
   surgery, does Aspirin usage, compared to no aspirin, cause bleeding" TYPE="CAUS">
      Aspirin does not increase bleeding or increase the need for allogeneic blood transfusion in coronary artery
      surgery.
   </CLAIM>
</REVIEW>
```

http://staffwww.dcs.shef.ac.uk/people/M.Stevenson/resources/bio_contradictions/

Predicting Claim Polarity

- Given question and claim, predict polarity
- SVM classifier
  - N-grams, negation terms, directionality terms (“less”), polarity terms (“alleviate’)

	F-score
Majority baseline	68.2
All features	87.3
Negation alone	83.3

Conclusion

- Multiple challenges in applying LBD
- Linguistic processing useful to reduce number of incorrect candidates that are generated
  - Syntactic analysis
  - Word sense disambiguation

Publications

- **Volume**
  - J. Preiss, M. Stevenson and R. Gaizauskas (2015) Exploring Relation Types for Literature-based Discovery. Journal of the American Medical Informatics Association.
- **Ambiguity**
  - J. Preiss and M. Stevenson (2016) The Effect of Word Sense Disambiguation Accuracy on Literature Based Discovery. BMC Medical Informatics and Decision Making
  - J. Preiss and M. Stevenson (2013) DALE: A Word Sense Disambiguation System for Biomedical Documents Trained using Automatically Labeled Examples. In *Proceedings of the 2013 NAACL HLT*, Atlanta, Georgia
  - M. Stevenson and Y. Guo (2010) Disambiguation of Ambiguous Biomedical Terms using Examples Generated from the UMLS Metathesaurus. *Journal of Biomedical Informatics*, 43(5): 762-773.
  - M. Stevenson, Y. Guo, R. Gaizauskas and D. Martinez (2008) Disambiguation of Biomedical Text using a Variety of Knowledge Sources. *BMC Bioinformatics*
- **Inconsistency**
  - A. Alamri and M. Stevenson (2016) A Corpus of Potentially Contradictory Research Claims from Cardiovascular Research Abstracts. *Journal of Biomedical Semantics*, 7(36).