Mining the History of Medicine

Sophia Ananiadou
National Centre for Text Mining
School of Computer Science
The University of Manchester
Outline

• Text Mining tools for enriching content

• Mining the History of Medicine

• Semantic annotations:
  – entities, events
  – OCR correction
  – Time-sensitive term inventory and search

• A prototype search system for the history of medicine
Mining the History of Medicine

• National Centre for Text Mining, School of Computer Science  www.nactem.ac.uk

• Centre for the History of Science, Technology and Medicine

AHRC
Digital Transformations
Big Data
Aim

Semantic search system
- British Medical Journal (BMJ) (1840 – present) (350K articles)
- London Medical Officer of Health reports (MoH) (1848 – 1972)

Text Mining: tools, resources, infrastructure

Provide different perspectives (e.g. medical, public health) on the treatment and prevention of diseases
- over time
- in different areas
Search historical archives

Keyword Search?

OCR

BMJ archive was digitised several years ago
- Word error rate can be over 30%
- Re-OCRing time-consuming and costly
- OCR errors problematic for TM methods

MOH OCRed more recently with better results

Keyword search over huge textual archives inefficient
Solutions

- Apply OCR improvement techniques
- Automatically detect **semantic** information in text
  - **Entities**: conditions, signs or symptoms, therapeutic measures
  - **Events**: symptoms caused by lung disease
- Automatically identify how the naming of concepts changes over time
Automatic query building, term suggestions
  • historical variants unknown to the user

Faceted search based on article metadata
  • Date, authors, title

Faceted search based on semantic content
  • Combinations of entities and/or relationships
OCR Post Correction

- Re-OCR BMJ too time consuming for current project
- Our solution: post-correct original OCR output based on spell-checker output

<table>
<thead>
<tr>
<th>SpellChecker</th>
<th>% errors resolved</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASpell</td>
<td>45%</td>
</tr>
<tr>
<td>Hunspell</td>
<td>60%</td>
</tr>
<tr>
<td>Word</td>
<td>63%</td>
</tr>
<tr>
<td>MAC OS</td>
<td>58%</td>
</tr>
</tbody>
</table>

Best open source spell checker
OCR Post Correction

- Hunspell produces ranked list of spelling corrections
- However, tailored to:
  - Human errors
  - General language

Our approach

**Augment** Hunspell dictionary with OpenMedSpel dictionary

**Override** default Hunspell correction ranking

- Rank suggestions according to *corpus frequency*
- Use *decade-specific word frequency* lists
JOHN MILLER, aged 26, glass-blower of Gateshead, was admitted August 27th, 1840, under the care of Sir John Fife, with the usual symptoms of stone in the bladder, which he has laboured under from childhood. At the age of 12 years he was admitted into this infirmary, but his friends would not allow the operation to be performed. He took a considerable quantity of medicine, which greatly relieved him. Since then the symptoms of stone have been constantly present, but have not been very urgent. A.- About a month ago, after bathing, the symptoms became much aggravated—so much so, as to compel him to apply again to the hospital for relief. On admission he was sallow and emaciated; laboured under symptoms of stone of a high degree, with great irritability of bladder, giving rise to a frequent desire to inicturate, and occasionally to the involuntary discharge of faces from straining. He passed a large quantity of mucus with his urine;—pulse small, tongue white, slight thirst, and appetite good.
JOHN MILLER, aged 26, glass-blower of Gateshead, was admitted August 27th, 1840, under the care of Sir John Fife, with the usual symptoms of stone in the bladder, which he has laboured under from childhood. At the age of 12 years he was admitted into this infirmary, friends world into allow the operation to be performed. Ike took a considerable (quantity of medicine, which greatly relieved him. Since then the symptom of stone have been constantly present, but have not been very urgent. A.- About a month ago, after bathing, the symptoms became much aggravated—so much so, as to compel him to apply again to the hospital for relief. On admission he was sallow and emaciated; laboured under symptoms of a high degree, with great irritability of bladder, giving rise to a frequent desire to inaccurate, and occasionally to the involuntary discharge of faces from straining. Ike passed a lag,e quantity of mucus with his urine; pulse small, tongue white, slight thirst, and appetite good.
JOHN MILLER, aged 26, glass-blower of Gateshead, was admitted August 27th, 1840, under the care of Sir John Fife, with the usual symptoms of stone in which he has laboured under from childhood. At the age of 12 years he was admitted into this infirmary, but his friends would not allow the operation to be performed. Lie took a considerable (quantity of medicine, which greatly relieved him. Since then the symptoms of stone have been constantly present, but have not been very urgent. About a month ago, after bathing, the symptoms aggravated—so much so, as to compel him to apply to the hospital for relief. On admission he was sallow and emaciated; laboured under symptoms of stone of a high degree, with great irritability of bladder, giving rise to a frequent desire to inaccurate, and occasionally to the involuntary discharge of faces from straining. Lie passed a large, e quantity of mucus with his urine;—pulse small, tongue white, slight thirst, and appetite good.
## Avoiding Over-Correction

<table>
<thead>
<tr>
<th>Original Text</th>
<th>Basic Hunspell</th>
<th>Frequency-driven Hunspell</th>
<th>Frequency-driven Hunspell augmented with OpenMedSpel</th>
</tr>
</thead>
<tbody>
<tr>
<td>antithrombotic</td>
<td>anthropometric</td>
<td>anthropometric</td>
<td>antithrombotic</td>
</tr>
<tr>
<td>pathophysiological</td>
<td>physiological</td>
<td>physiological</td>
<td>pathophysiological</td>
</tr>
<tr>
<td>thromboembelism</td>
<td>thrombosis</td>
<td>thrombosis</td>
<td>thromboembolism</td>
</tr>
<tr>
<td>warfarin</td>
<td>wayfaring</td>
<td>warfarin</td>
<td>warfarin</td>
</tr>
<tr>
<td>transoesophageal</td>
<td>oesophageal</td>
<td>oesophageal</td>
<td>transoesophageal</td>
</tr>
<tr>
<td>echocardiography</td>
<td>electroocardiography</td>
<td>oesophageal</td>
<td>echocardiography</td>
</tr>
<tr>
<td>intracardiac</td>
<td>intra cardiac</td>
<td>intracellular</td>
<td>intracardiac</td>
</tr>
<tr>
<td>tomogram</td>
<td>mammogram</td>
<td>tomogram</td>
<td>tomogram</td>
</tr>
</tbody>
</table>
✓ Gold-standard collection of 24 hand-corrected documents
  ✓ 3 documents from: 1840s, 1860s, 1880s, 1900s, 1920s, 1940s, 1960s, 1980s
✓ Documents selected according to likelihood containing OCR errors
## Evaluation (Word Level Accuracy)

<table>
<thead>
<tr>
<th>PMCID</th>
<th>Raw OCR</th>
<th>Basic Hunspell</th>
<th>Frequency-driven Hunspell</th>
<th>Basic Hunspell with OpenMedSpel</th>
<th>Frequency-driven Hunspell + OpenMedSpel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1840s</td>
<td>75.9%</td>
<td>83.6%</td>
<td>84.9%</td>
<td>83.3%</td>
<td>85.6%</td>
</tr>
<tr>
<td>1860s</td>
<td>86.6%</td>
<td>90.0%</td>
<td>91.2%</td>
<td>90.2</td>
<td>93.1%</td>
</tr>
<tr>
<td>1880s</td>
<td>82.6%</td>
<td>87.4%</td>
<td>88.9%</td>
<td>88.0%</td>
<td>89.0%</td>
</tr>
<tr>
<td>1900s</td>
<td>70.2%</td>
<td>80.7%</td>
<td>82.3%</td>
<td>81.7%</td>
<td>84.3%</td>
</tr>
<tr>
<td>1920s</td>
<td>74.2%</td>
<td>79.9%</td>
<td>82.7%</td>
<td>81.6%</td>
<td>84.3%</td>
</tr>
<tr>
<td>1940s</td>
<td>86.6%</td>
<td>88.1%</td>
<td>89.5%</td>
<td>89.4%</td>
<td>91.3%</td>
</tr>
<tr>
<td>1960s</td>
<td>94.9%</td>
<td><strong>91.9%</strong></td>
<td><strong>93.5%</strong></td>
<td>95.1%</td>
<td><strong>95.7%</strong></td>
</tr>
<tr>
<td>1980s</td>
<td>93.0%</td>
<td>87.2%</td>
<td>89.9%</td>
<td><strong>92.6%</strong></td>
<td>94.8%</td>
</tr>
</tbody>
</table>
Rule-based OCR correction

- Many regular OCR character recognition errors occur in BMJ
  - Two characters being recognised instead of one,
    - e.g. $h \rightarrow lh$, $t \rightarrow lt$, $m \rightarrow mn$, $u \rightarrow ui$
  - Punctuation characters/digits erroneously appearing in words
    - e.g. $large$

- Many character insertion problems corrected by Hunspell
  - Still some regular errors remain uncorrected

- Solution – apply rule-based pre-processing prior to Hunspell
Rule-based OCR correction

- Eight pattern replacement rules, i.e., $tll \rightarrow th$, $wl \rightarrow w$, $lh \rightarrow h$, $hl \rightarrow h$, $mn \rightarrow m$, $nm \rightarrow m$, $wv \rightarrow w$ and $ii \rightarrow l$
- Remove punctuation characters and digits appearing in words, except when:
  - Digits are followed by letters denoting ordinal numbers not removed
  - Apostrophes denoting possessives not removed
  - Hyphens are only removed when they occur as the second or penultimate character of a word.
    - Hyphens in other positions are likely to form part of valid hyphenated words
<table>
<thead>
<tr>
<th>PMCID</th>
<th>Frequency-driven Hunspell + OpenMedSpel</th>
<th>Frequency-driven Hunspell + OpenMedSpel + Rule-based pre-processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1840s</td>
<td>85.6%</td>
<td>87.0%</td>
</tr>
<tr>
<td>1860s</td>
<td>93.1%</td>
<td>94.1%</td>
</tr>
<tr>
<td>1880s</td>
<td>89.0%</td>
<td>90.5%</td>
</tr>
<tr>
<td>1900s</td>
<td>84.3%</td>
<td>86.2%</td>
</tr>
<tr>
<td>1920s</td>
<td>84.3%</td>
<td>87.6%</td>
</tr>
<tr>
<td>1940s</td>
<td>91.3%</td>
<td>92.3%</td>
</tr>
<tr>
<td>1960s</td>
<td>95.7%</td>
<td>94.2%</td>
</tr>
<tr>
<td>1980s</td>
<td>94.8%</td>
<td>92.7%</td>
</tr>
</tbody>
</table>
Extracting semantic types, events

- Supervised methods
- Corpus composition
  - 25 articles from BMJ
  - 4 extracts from MOH
  - 70,000 words
- Articles from 4 key decades
  - 1850s, 1890s, 1920 and 1960s
  - Capture language/terminology changes over time
## HIMERA corpus: Entity Types

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatomical</td>
<td>An entity forming part of the human body</td>
<td>lung, lobe, sputum, fibroid</td>
</tr>
<tr>
<td>Biological Entity</td>
<td>A living entity not part of the human body</td>
<td>tubercle bacilli, mould, guinea-pig, flea</td>
</tr>
<tr>
<td>Condition</td>
<td>Medical condition or ailment</td>
<td>phthisis, bronchitis, typhoid</td>
</tr>
<tr>
<td>Environmental</td>
<td>Environmental factor relevant to incidence/prevention/control/treatment of condition</td>
<td>humidity, high mountain climates, milk, linen, drains</td>
</tr>
<tr>
<td>Sign_or_Symptom</td>
<td>Altered physical appearance or behaviour as a probable result of injury or condition</td>
<td>cough, pain, rise in temperature, swollen</td>
</tr>
<tr>
<td>Subject</td>
<td>Individual or group of cases under discussion</td>
<td>asthma patients, those with negative reactions to tuberculin</td>
</tr>
<tr>
<td>Therapeutic_or_Investigational</td>
<td>Treatment, substance, medium or procedure, prescribed or used in investigation</td>
<td>atrophine sulphate, generous diet, change of air, lobectomy</td>
</tr>
</tbody>
</table>
## HIMERA Corpus: Events

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affect</td>
<td>A entity or event is affected, infected, changed or transformed, possibly by another entity or event</td>
<td>Cause: The cause of the affection Target: The entity or event affected Subject: The medical subject affected</td>
</tr>
</tbody>
</table>

- **Anatomical**: the fluid effusion is limited by adhesions.
- **Therapeutic_or_Investigational**: The patient did well with only a few doses of ammonia.
- **Environmental**: these diseases depend upon their surroundings or environments.
- **Environmental**: A stagnant ditch of decomposing miasmatic matter, which must be injurious to the inhabitants.
- **Therapeutic_or_Investigational**: A few leeches also may afford much relief to the pain.
<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Participants</th>
</tr>
</thead>
</table>
| Cause | An entity or event results in the manifestation of another entity or event | Cause: The cause of the event  
Result: The resulting entity or event  
Subject: The medical subject affected |

I believe the pain was seated in the pillars of the diaphragm, as it was chiefly occasioned by hiccough. The vast majority of cases of phthisis are of tuberculous origin.

In one form of the disease, viz., abdominal tuberculosis, the great bulk of the cases may be regarded as due to ingestion of tuberculous milk. Still-birth and premature birth are due to infection of the mother.

Game decayed only sufficiently to please the taste of the epicure, has produced severe cholera in persons not accustomed to eat it in that state.
## Entity Agreement Rates (F-Score)

<table>
<thead>
<tr>
<th>Type</th>
<th>Exact span agreement</th>
<th>Relaxed span agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatomical</td>
<td>0.81</td>
<td>0.85</td>
</tr>
<tr>
<td>Biological_Entity</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Condition</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td>Environmental</td>
<td>0.63</td>
<td>0.79</td>
</tr>
<tr>
<td>Sign_or_Symptom</td>
<td>0.84</td>
<td>0.88</td>
</tr>
<tr>
<td>Subject</td>
<td>0.70</td>
<td>0.81</td>
</tr>
<tr>
<td>Therapeutic_or_Investigational</td>
<td>0.73</td>
<td>0.78</td>
</tr>
<tr>
<td>TOTAL</td>
<td>0.80</td>
<td>0.86</td>
</tr>
<tr>
<td>Entity Type</td>
<td>Count</td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>Anatomical</td>
<td>2002</td>
<td></td>
</tr>
<tr>
<td>Biological_Entity</td>
<td>295</td>
<td></td>
</tr>
<tr>
<td>Condition</td>
<td>1499</td>
<td></td>
</tr>
<tr>
<td>Environmental</td>
<td>1268</td>
<td></td>
</tr>
<tr>
<td>Sign_or_Symptom</td>
<td>1171</td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td>1062</td>
<td></td>
</tr>
<tr>
<td>Therapeutic_or_Investigational</td>
<td>1046</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>8343</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affect</td>
<td>611</td>
</tr>
<tr>
<td>Causality</td>
<td>204</td>
</tr>
<tr>
<td>TOTAL</td>
<td>815</td>
</tr>
</tbody>
</table>

- Discrepancies in double annotated docs resolved
- *Anatomical entities* most numerous
- *Biological_Entity* sparse according to rare mention of microorganisms prior to 20th century
Entity Models trained NERSuite
nersuite.nlplab.org
  • Pre-processing using GENIA tagger to generate features
✓ Default features for learning
  • Surface word forms
  • Base word forms
  • POS tags
  • Syntactic chunk tags
✓ Default features augmented with semantic features
Entity recognition Models

- **Baseline (BL)** – Default NERSuite features
- **Full MetaMap (FM)** – Semantic features added by applying MetaMap semantic tagger (133 types)
- **Selective MetaMap (SM)** – Semantic features added by mapping selected MetaMap categories to the seven HYMERA entity categories
- **UMLS Lookup (UL)** – Semantic features assigned using dictionary lookup in filtered version of the UMLS Metathesaurus dictionary.
## MetaMap Categories

<table>
<thead>
<tr>
<th>MetaMap Categories</th>
<th>Hymera category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatomical Abnormality, Body Substance   Body Part, Organ, or Organ Component</td>
<td>Anatomical</td>
</tr>
<tr>
<td>Body Location or Region, Body Space or Junction, Tissue</td>
<td></td>
</tr>
<tr>
<td>Animal, Mammal, Cell, Bacterium, Organism</td>
<td>Biological_ Entity</td>
</tr>
<tr>
<td>Food, Chemical Viewed Structurally Element, Ion, or Isotope, Hazardous or Poisonous</td>
<td>Environmental</td>
</tr>
<tr>
<td>Substance, Natural Phenomenon or Process</td>
<td></td>
</tr>
<tr>
<td>Disease or Syndrome, Pathologic Function</td>
<td>Condition</td>
</tr>
<tr>
<td>Clinical Drug, Amino Acid, Peptide, or Protein, Immunologic Factor, Organic</td>
<td>Therapeutic_or_</td>
</tr>
<tr>
<td>Chemical, Pharmacologic Substance, Biologically Active Substance, Lipid</td>
<td>Investigational</td>
</tr>
<tr>
<td>Sign or Symptom, Finding</td>
<td>Sign_or_ Symptom</td>
</tr>
<tr>
<td>Group, Patient or Disabled Group</td>
<td>Subject</td>
</tr>
</tbody>
</table>
NERSuite Results (F-Score, 5-fold cross validation)

<table>
<thead>
<tr>
<th>Type</th>
<th>BL (Exact)</th>
<th>BL (Relaxed)</th>
<th>UL (Exact)</th>
<th>UL (Relaxed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatomical</td>
<td>0.73</td>
<td>0.79</td>
<td>0.76</td>
<td>0.83</td>
</tr>
<tr>
<td>Biological_Entity</td>
<td>0.70</td>
<td>0.72</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>Condition</td>
<td>0.73</td>
<td>0.82</td>
<td>0.77</td>
<td>0.86</td>
</tr>
<tr>
<td>Environmental</td>
<td>0.49</td>
<td>0.58</td>
<td>0.49</td>
<td>0.60</td>
</tr>
<tr>
<td>Sign_or_Symptom</td>
<td>0.66</td>
<td>0.71</td>
<td>0.69</td>
<td>0.74</td>
</tr>
<tr>
<td>Subject</td>
<td>0.74</td>
<td>0.81</td>
<td>0.75</td>
<td>0.81</td>
</tr>
<tr>
<td>Therapeutic_or_Investigational</td>
<td>0.57</td>
<td>0.64</td>
<td>0.61</td>
<td>0.67</td>
</tr>
<tr>
<td>ALL</td>
<td>0.67</td>
<td>0.74</td>
<td>0.70</td>
<td>0.77</td>
</tr>
</tbody>
</table>

UL results reported as most straightforward model to apply MetaMap can be slow
Comparison

- **Anatomical** F-Score 0.82
- Comparable with anatomical entity recognition for (0.85 F-Score) (Pyysalo & Ananiadou, 2014)
- **Condition** F-Score of 0.76 (exact) and 0.86 relaxed
  - Comparable to disease recognition performance for electronic health records (0.75 exact/0.88 relaxed) (Kaewphan et al., 2014)


How time-sensitive is NER?

Split the HIMERA to decades
- Used docs from a single decades as test data
- Trained on combinations of docs from other decades
Time-sensitive NER example
(Testing on 1890s docs)

<table>
<thead>
<tr>
<th>Type</th>
<th>1850s</th>
<th>1920s</th>
<th>1960s</th>
<th>1850s/1920s</th>
<th>1920s/1960s</th>
<th>1850s/1920s/1960s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatomical</td>
<td>0.78</td>
<td>0.66</td>
<td>0.53</td>
<td>0.83</td>
<td>0.66</td>
<td>0.84</td>
</tr>
<tr>
<td>Biological Entity</td>
<td>0.41</td>
<td>0.87</td>
<td>0.13</td>
<td>0.88</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td>Condition</td>
<td>0.79</td>
<td>0.77</td>
<td>0.72</td>
<td>0.85</td>
<td>0.67</td>
<td>0.85</td>
</tr>
<tr>
<td>Environmental</td>
<td>0.32</td>
<td>0.26</td>
<td>0.23</td>
<td>0.38</td>
<td>0.29</td>
<td>0.44</td>
</tr>
<tr>
<td>Sign or Symptom</td>
<td>0.68</td>
<td>0.56</td>
<td>0.41</td>
<td>0.72</td>
<td>0.54</td>
<td>0.72</td>
</tr>
<tr>
<td>Subject</td>
<td>0.79</td>
<td>0.73</td>
<td>0.68</td>
<td>0.80</td>
<td>0.72</td>
<td>0.78</td>
</tr>
<tr>
<td>Therapeutic or Investigational</td>
<td>0.35</td>
<td>0.39</td>
<td>0.19</td>
<td>0.47</td>
<td>0.33</td>
<td>0.52</td>
</tr>
<tr>
<td>ALL</td>
<td>0.70</td>
<td>0.64</td>
<td>0.51</td>
<td>0.76</td>
<td>0.61</td>
<td>0.77</td>
</tr>
</tbody>
</table>
Observations

- Similar pattern when different decades are test set
- When single decade used for training
  - Best results normally when decade immediately prior to test data is used for training
- When two decades used for training
  - Best results usually when decades surrounding test decade are used for training
- Best overall results when 3 decades used for training
  - But, usually only small performance difference from two surrounding decades
- **Temporal closeness** of training and test data more important than volume of training data
- As long as some temporally close training data is included, it is not harmful to include training data from more distant decades
Event Recognition

- EventMine [www.nactem.ac.uk/EventMine/](http://www.nactem.ac.uk/EventMine/)
- Detects triggers: origin
- Finds individual event participants
- Combines trigger-argument pairs into semantic structures
Event trigger recognition performance

<table>
<thead>
<tr>
<th>Event Type</th>
<th>F-Score (Exact)</th>
<th>F-Score (Relaxed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affect</td>
<td>0.40</td>
<td>0.45</td>
</tr>
<tr>
<td>Causality</td>
<td>0.19</td>
<td>0.27</td>
</tr>
<tr>
<td>TOTAL</td>
<td>0.35</td>
<td>0.42</td>
</tr>
</tbody>
</table>

- Much sparser data than previous EventMine applications
- Event types have wide semantic scope
- Diverse range of triggers and stylistic variations can make recognition a challenge
- Events with multiple participants recognised
✓ Ongoing work to triple the size of manually annotated corpus
  • Different decades to provide greater evidence of language usage
  • More events
History of Medicine Terms

- Concepts can be expressed in many ways
- We take into account such variants to improve search

<table>
<thead>
<tr>
<th>Lexical Variations</th>
</tr>
</thead>
<tbody>
<tr>
<td>oedema</td>
</tr>
<tr>
<td>whooping cough</td>
</tr>
<tr>
<td>pulmonary tuberculosis</td>
</tr>
<tr>
<td>respiratory diseases</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Semantic Variations</th>
</tr>
</thead>
<tbody>
<tr>
<td>tuberculosis</td>
</tr>
<tr>
<td>smallpox</td>
</tr>
</tbody>
</table>
Why not using existing resources?

- MeSH, Disease Ontology, UMLS, Metathesaurus

Used in search systems to *expand query terms*

Textual variants of concepts *change over time*

- Existing resources tend to focus on *contemporary* terminology
- *Historically-relevant* variants not covered in a comprehensive and consistent manner
Term Variation in BMJ

Phthisis vs. Tuberculosis

- Phthisis
- Tuberculosis

Graph showing the variation of terms 'Phthisis' and 'Tuberculosis' in BMJ from 1840 to 1983.
Term Variation in BMJ
*Scarlet Fever vs. Scarlatina*

![Graph showing term variation over years with data points for Scarlet Fever and Scarlatina.]
Time Sensitive Terminological Inventory

✓ Time sensitive inventory of medical terminology
  • Complements existing terminological resources
  • Accounts for term variants over time

✓ How?
  • Extracted term variants from 19th c. medical lexica
  • Identified more variants using *distributional semantics* to BMJ and MOH archives
## 19th century medical lexica: NLM Archive

<table>
<thead>
<tr>
<th>Name</th>
<th>Year</th>
<th># top-level entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nomenclature of diseases</td>
<td>1872</td>
<td>1347</td>
</tr>
<tr>
<td>Index of diseases, their symptoms, and treatment</td>
<td>1882</td>
<td>522</td>
</tr>
</tbody>
</table>
• English main term
• Latin main term
• Synonym list
• Sub-type English term
• Sub-type Latin term

Processing Historical Lexica - example

68. Osseous tumor, or bony tumor. Tumor osseus.
     Ivory-like. Eberneus.
     Cancellated. Cancellatus.
   • Osseous tumor of the soft parts. Tumor osseus partium mollium.

   • Fibro-cartilaginous tumor. Tumor fibro-cartilaginosus.

    A tumor growing in or near a gland, and more or less perfectly resembling it in structure.
   • Lymphoid tumor. Lymphoma.
     Composed of corpuscles resembling the cell-elements of lymphatic glands.

   • Aneurism by anastomosis, or Racemose aneurism. Aneurysma racemosum.
   • Nævus. Nævus.
Processing Historical Lexica

✓ Link UMLS concepts to historical concepts
  • approximate string matching
✓ 1,588 UMLS and historical concepts linked
  • 2,422 historical synonyms/variants not in UMLS
    • e.g., *cerebro-spinal fever* as a synonym of *epidemic meningitis only* in historical resource
✓ 800 historical concepts could not be linked to UMLS concepts
✓ *lack of comprehensive historical coverage in UMLS*
Need for Distributional Semantics

- Different lexica to account for terminology usage in different time periods
- Distributional semantics techniques –
- Processed complete BMJ and MOH archives to extract semantically related terminology across complete archive time-span
TM Methods for Historical Variant Identification

- Pre-processing
- Distributional similarity measures
- Detection of historically relevant term variants
Extracted disease terms

Combined 2 disease NERs

Dictionary matching from 19th lexica
  ✓ Index of Diseases
  ✓ Nomenclature of Diseases

Trained on NCBI Disease Corpus
Distributional Similarity Measures

- Generate Context vector for each term
- 6 word window around term
- High cosine similarity between context vectors indicates related terms

![Distributional Similarity Measures Table]

\[
\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = 0.8434
\]
Distributional Similarity Measures

scarlet fever

scarlatina 0.8434
whooping cough 0.7415
measles 0.7094
diphtheria 0.6689
acute pneumonia 0.6501
chicken pox 0.6407
german measles 0.6336
diarrhoeal disease 0.6301
infantile diarrhoea 0.6234
diarrhea 0.6208
croup 0.6093
mump 0.6042
typhus fever 0.5535
...

Automatic Evaluation of DSM output

- Use UMLS Metathesaurus
  - At least one UMLS synonym is identified >70% cases
  - MOH reports, *all* synonyms listed in UMLS discovered by the DSM method in 49% of cases
Expert Evaluation of DSM output

- Automatic evaluation general performance of DSMs
- However
  - No historical synonym coverage
  - No other types of semantic relations between related terms
- Expert evaluation covers
  - 348 source-target term pairs examined by expert, with cosine similarity greater than 0.8
  - Only 18% identified as synonyms using UMLS
  - Expert asked to characterise semantic relation (if any) between the two terms
Possible relations between terms

- isA
- isParent
- Synonym
- Spatially related
- Other
- Unrelated

Affects
isAffectedBy
### Expert evaluation results

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonym</td>
<td>83 (24%)</td>
</tr>
<tr>
<td>isA</td>
<td>21 (6%)</td>
</tr>
<tr>
<td>isParent</td>
<td>35 (10%)</td>
</tr>
<tr>
<td>Affects</td>
<td>10 (3%)</td>
</tr>
<tr>
<td>isAffectedBy</td>
<td>11 (3%)</td>
</tr>
<tr>
<td>Spatially Related</td>
<td>19 (5%)</td>
</tr>
<tr>
<td>OtherRelation</td>
<td>35 (10%)</td>
</tr>
<tr>
<td><strong>TOTAL SEMANTICALLY RELATED</strong></td>
<td><strong>216 (62%)</strong></td>
</tr>
<tr>
<td>Unrelated</td>
<td>106 (30%)</td>
</tr>
<tr>
<td>Non-disease pair</td>
<td>26 (7%)</td>
</tr>
</tbody>
</table>

- Expert identified more synonyms than UMLS
- DSM can identify more synonyms
- 62% of pairs semantically related
Semantic annotation of archives

- Integrated: OCR correction, NER, event into interoperable pipeline for medical history
- Automatically recognised NEs used as input terms to DSM methods
  - Recognised semantic relations between a large number of medical terms belonging to different semantic classes
  - Created a wide coverage, historically-aware resource of medical terminology
Access “query builder” to start search refinement.
User can refine search using multiple criteria. Here we start by searching for a specific term.
User enters term of interest and then clicks on “Refine Search”
User selects “pulmonary tuberculosis”

Current search criteria

Frequency of “pulmonary consumption” over time

Retrieved documents

related terms
“pulmonary tuberculosis” is added to query builder

Terms related to “pulmonary consumption” and “pulmonary tuberculosis”
Further refine retrieved documents by date published.
Only retain documents published up until 1950
Further refine retrieved documents by date published.
Refine by automatically recognised semantic entities
Semantic refinement can be made according to different entity categories.
Here “Environmental” entities present in the current document set, with their counts. The user may type into the box to narrow the list of entities displayed.
New semantic restriction added to query builder display

Number of documents further reduced
Results can also be refined by searching for documents containing involving entities
User can choose between the 2 events recognised by our system.
User (partially) completes template according to the types of events they are searching for. Here, the search is for documents mentioning causes of tuberculosis.
Pie chart shows the distribution of entity types in a document. Hovering the mouse over a section reveals the entity type in question.
THE attention.

370 Medical JOURNAL I

INTERNATIONAL MEDICAL CONGRESS:

[AUG. 16, 1913.]

THE NECESSITY FOR A MORE THOROUGH

CONTROL OF THE MILK SUPPLY IN COMBATING SURGICAL TUBERCULOSIS

IN CHILDREN.

By H. Delepine, M.D.

SURGERY DEPARTMENT.

THE results of the investigations by Delepine and others, have done valuable work in developing the methods by means of which the tuberculous milk which is supplied to our large cities.

Professor Delepine has for several years carried out anti-antiport investigations in Manchester in regard to the tuberculous milk question, and he has shown that much still remains to be done to diminish the amount of tuberculous milk which is supplied to our large cities.

The control of the Edinburgh milk supply is very inadequate, and it is a remarkable fact that the Public Health Department has no power to enforce the slaughter of the cows with tuberculosis. Most tuberculous milk is supplied to Edinburgh from time to time in the city by other localities; the medical officer of health has merely power to enforce the boundaries.

Fortunately, comparatively large quantities of the bacilli are necessary to produce tuberculosis, even in more prevalent amongst children than it is. Post-mortem records in this country show that at about 25 per cent. of the children possessing tuberculous lesions the infection has taken place through the alimentary tract.

As the result of my clinical experience in the Royal Edinburgh Hospital for Sick Children, I have for many years been convinced that Koch and his disciples commit a serious error in practically disregarding the milk as a source of tuberculous infection in children. I could adduce many instances from my own personal observation where the etiological relations between the disease and the milk history of the

Number of "Environmental" entities

"Environmental" entities are highlighted

"Environmental" entity instances in the document
Cases where specific causes of tuberculosis are mentioned in the document

Phrases identifying Causality events are highlighted
Affect events allow us to explore, e.g., whether therapeutic measures have a positive or negative effect.
Conclusions

- Explore further archives using topic analysis
- Automatic clustering of search results according to similarities of semantic content
- Expand types of entities and events and enrich further BMJ
- Link with other archives
- Create a rich annotation database
Acknowledgments

• BMJ and Wellcome Library
• AHRC
• NaCTeM and CHSTM
• Paul Thompson, Elizabeth Toon, Jacob Carter, George Kontonatsios, Riza Batista-Navarro

Investigators: S. Ananiadou, J. McNaught (NaCTeM)
C. Timmermann, M. Worboys (CHTM)