Big Data Challenges with Big Corpora and Big Taxonomies

Dr Paul Rayson
Director of UCREL research centre, School of Computing and Communications, Lancaster, UK

Joint work with Alistair Baron, Scott Piao & Steve Wattam
Larger and larger corpora...

- **Brown/LOB**
  - 1960s
  - 1 million

- **BNC**
  - 1990s
  - 100 million

- **Web**
  - Social Media
  - Digitisation
  - Present day
  - ? billions
A (brief) History of Computerised Corpus Tools (Mura Nava)

http://timemapper.okfnlabs.org/muranava/history-of-computerised-corpus-tools
A (brief) History of Computerised Corpus Tools (Mura Nava)

http://timemapper.okfnlabs.org/muranava/history-of-computerised-corpus-tools
Tool-driven linguistics?


• Workshop at CL2015, Lancaster, July 20-24
  – Brainstorming the next generation of corpus software
  – Organizers: Laurence Anthony, Stephen Wattam, Matt Coole, John Mariani, Paul Rayson, John Vidler
THE HISTORY OF THE LATE CONSPIRACY AGAINST THE KING AND THE NATION.

With a Particular Account of the LANCASHIRE PLOT, and All the other Attempts and Machinations of the disaffected Party, since His Majesty's Accession to the Throne.

Extracted out of the Original Informations of the Witnesses, and other Authentick Papers.

LONDON,
Printed for Daniel Brown, at the Black Swan and Bible without Temple-Bar, and Tho. Bennet at the Half-Moon in St. Paul's Church-yard. MDC XCVI.
Though I **speake** with the tongues of men & of Angels, and **haue** not **charity**, I am become as sounding **brasse** or a tinkling cymbal. And though I **haue** the gift of **prophesie**, and **vnderstand** all mysteries and all knowledge: and though I **haue** all faith, so that I could **re moue mountaines**, and **haue** no **charitie**, I am nothing...

*(Authorised Version of the Bible, 1611)*
SAMUELS project

http://www.gla.ac.uk/samuels/

- SAMUELS: Semantic Annotation and Mark-Up for Enhancing Lexical Searches
  - funded by the Arts and Humanities Research Council in conjunction with the Economic and Social Research Council (grant reference AH/L010062/1)
  - January 2014 to April 2015
- Aim
  - deliver a system for automatically annotating words in texts with their precise meanings, disambiguating between possible meanings of the same word
  - will provide for each word in a text the Historical Thesaurus of English reference code for that concept.
- Project team:
  - Lancaster: Alistair Baron, Scott Piao, Steve Wattam
  - University of Glasgow (lead institution), Lancaster University, University of Huddersfield, University of Central Lancashire, University of Strathclyde, Oxford University Press
  - international partners: Brigham Young University (Utah), Åbo Akademi University (Finland), and the University of Oulu (Finland).
Big Data Challenges

• Big corpora:
  – Early English Books Online (EEBO) Text Creation Partnership (TCP) consisting of over 53,830 books published between 1473 and 1700 (1.27 billion words; Phase 2 November 2014 release)
  – Two hundred years of UK Parliamentary Hansard consisting of over 7 million files (~2 billion words)

• Big taxonomies:
  – Historical Thesaurus of English (developed at the University of Glasgow) and the Oxford English Dictionary to help us improve methods for the automatic semantic analysis of historical texts.
  – The Historical Thesaurus contains 793,742 word forms arranged into 225,131 semantic categories.
Big Data Challenges

• The combination of scale (and historical nature) of the corpora and the taxonomy pose significant computational challenges for existing retrieval methods (Wmatrix) and annotation software (USAS)

• Our solutions
  – Variant Spelling methods
  – Improved semantic disambiguation techniques (Historical Thesaurus Semantic Tagger – HTST)
  – Use of big data methods e.g. cluster and cloud computing
• Addition or removal of ‘e’, e.g. aske, workes, dos
• Doubling and singling of letters, e.g. smels, heere, leggs
• Interchanged letters: \{ u , v \}, \{ j , i \}, \{ ie , y \}, \{ vv , w \}, e.g. haue, vnder, maiestie, vvas
• Usage of apostrophe, e.g. vow’d, ‘em
• Spellings which are variable still today, e.g. centre/center, -or/-our, -ise/-ize
• Fused forms, e.g. t’is, t’was, o’th
• Archaic –(e)th and –(e)st endings, e.g. hath, doth, seemeth, shouldst
• Archaic forms, e.g. betwixt, howbeit
• Phonetic spellings, e.g. publiquely, blew (blue)
• + any combination of the above and other irregular spellings, e.g. ligge (Jig), diuell (devil), shak’d (shook)
The extent of spelling variation in EmodE corpora

- And its effect on corpus methods such as keywords
% Variant Types

Decade

ARCHER
EEBO
Innsbruck
Lampeter
EMEMT
Shakespeare
Average Trend
• Searching for words can be problematic: *would, wolde, woolde, wuld, wulde, wud, wald, vvould, vvold*, etc.

• Frequencies split by multiple spellings.

• Knock-on effect on key words (Baron *et al.*, 2009), key word clusters (Palander-Collin & Hakala, 2011) and collocates.
The need for normalisation ...

- Automatic semantic analysis of EmodE corpora

- Automatic POS tagging of historical corpora

- Corpus annotation in general
With no standardization
After automatic standardization
After manual standardization
When applied to Modern British English

CLAWS part of speech tagging accuracy (%)

Shakespeare
Lampeter

Corpus sample
VARD (VARiant Detector)

http://ucrel.lancs.ac.uk/vard/
• Freely available for academic use: http://ucrel.lancs.ac.uk/vard/

• Designed to assist researchers in standardising spelling variation in historical corpora both manually and automatically.

• Uses methods from modern spellchecking to find spelling variants and offer/select appropriate modern equivalents.

• The original spelling is always retained in the text with an xml tag surrounding the replacement.
  
  – <normalised orig="charitie">charity</normalised>

• Allows for the use of standard corpus linguistics tools without any modification.

• Used to normalise released historical (and other) corpora, e.g. EMEMT (Lehto et al., 2010) and CEEC (Palander-Collin & Hakala, 2011).
Though I *speake* with the tongues of men & of Angels, and *haue* not charity, I am become as sounding *brasse* or a tinkling cymbal. And though I *haue* the gift of *prophesie*, and *vnderstand* all mysteries and all knowledge: and though I *haue* all faith, so that I could *remoue* *mountaines*, and *haue* no *charitie*, I am nothing...

*(Authorised Version of the Bible, 1611)*
USAS (Modern English) semantic tagger

- Full text tagging, not just selected words (c.f. Diction, LIWC, RID)
- Tagging the coarse-grained sense in context, not just the word
- Not task specific categories
- Flexible category set with hierarchical structure
- Words and multi-word expressions (MWE) e.g. phrasal verbs (stubbed out), noun phrases (riding boots), proper names (United States of America), true idioms (living the life of Riley)
<table>
<thead>
<tr>
<th>Column</th>
<th>Row</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>General and abstract terms</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>The body and the individual</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>Arts and crafts</td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td>Emotion</td>
</tr>
<tr>
<td>F</td>
<td>5</td>
<td>Food and farming</td>
</tr>
<tr>
<td>G</td>
<td>6</td>
<td>Government and public</td>
</tr>
<tr>
<td>H</td>
<td>7</td>
<td>Architecture, housing and the home</td>
</tr>
<tr>
<td>I</td>
<td>8</td>
<td>Money and commerce in industry</td>
</tr>
<tr>
<td>K</td>
<td>9</td>
<td>Entertainment, sports and games</td>
</tr>
<tr>
<td>L</td>
<td>10</td>
<td>Life and living things</td>
</tr>
<tr>
<td>M</td>
<td>11</td>
<td>Movement, location, travel and transport</td>
</tr>
<tr>
<td>N</td>
<td>12</td>
<td>Numbers and measurement</td>
</tr>
<tr>
<td>O</td>
<td>13</td>
<td>Substances, materials, objects and equipment</td>
</tr>
<tr>
<td>P</td>
<td>14</td>
<td>Education</td>
</tr>
<tr>
<td>Q</td>
<td>15</td>
<td>Language and communication</td>
</tr>
<tr>
<td>S</td>
<td>16</td>
<td>Social actions, states and processes</td>
</tr>
<tr>
<td>T</td>
<td>17</td>
<td>Time</td>
</tr>
<tr>
<td>W</td>
<td>18</td>
<td>World and environment</td>
</tr>
<tr>
<td>X</td>
<td>19</td>
<td>Psychological actions, states and processes</td>
</tr>
<tr>
<td>Y</td>
<td>20</td>
<td>Science and technology</td>
</tr>
<tr>
<td>Z</td>
<td>21</td>
<td>Names and grammar</td>
</tr>
</tbody>
</table>
Lexical resources

- Lexicon of 56,316 items
  - presentation NN1 Q2.2 A8 S1.1.1 K4
- MWE list of 18,971 items
  - travel_NN1 card*_NN* M3/Q1.2
- A small wildcard lexicon
  - *kg NNU N3.5
- Unknown words using WordNet synonym lookup
Disambiguation methods (1)

1. POS tag
   - *spring* noun  [season sense] [coil sense]
   - *spring* verb  [jump sense]

2. General likelihood ranking for single-word and MWE tags
   - *green* referring to [colour] is generally more frequent than *green* meaning [inexperienced]

3. Overlapping MWE resolution
   - Heuristics applied: semantic MWEs override single word tagging, length and span of MWE also significant
Disambiguation methods (2)

• 4. Domain of discourse
  – adjective *battered*
    • [Violence] (e.g. battered wife)
    • [Judgement of Appearance] (e.g. battered car)
    • [Food] (e.g. battered cod)

• 5. Text-based disambiguation
  – one sense per text

• 6. Template rules
  – *Auxiliary verbs* (*be/do/have*)
  – *account* of NP [narrative]
  – balance of xxx *account* [financial]
Disambiguation methods (3)

- 7. Local probabilistic
  - *account* occurring in the company of *financial, bank, overdrawn, money*
  - surrounding words, POS tags or semantic fields
  - span of words
  - co-occurrence measures rather than HMM
Evaluation (modern data)

- Hand tagged test corpus of 124,839 words
- Error rate of 8.95%
- Ambiguity ratio 47.73%
- Reduced to 17.06% by disambiguation
- Not all ambiguity is resolved, but 1st choice tag selection gives 91% accuracy.
I The external world
01 The world
  01.01 The earth
  01.01.01 Region of the earth
  01.01.02 Geodetic references
  01.01.03 Direction
  01.01.04 Land
  01.01.05 Water
  01.01.06 Named regions of earth
  01.01.07 Structure of the earth
  01.01.08 Minerals
  01.01.09 Earth science
  01.01.10 The universe
  01.01.11 Atmosphere, weather
  01.02 Life
  01.02.01 Health and disease
  01.02.02 Death
  01.02.03 Biology
  01.02.04 Plants
  01.02.05 The body
  01.02.06 Animals
  01.02.07 People
  01.02.08 Food and drink
  01.02.09 Textiles

II The mental world
02 The mind
  02.01 Mental capacity
  02.01.01 Spirituality
  02.01.02 Intellect
  02.01.03 Consciousness
  02.01.04 Disposition/character
  02.01.05 The psyche
  02.01.06 Thought
  02.01.07 Perception/cognition
  02.01.08 Understanding
  02.01.09 Lack of understanding
  02.01.10 Intelligibility
  02.01.11 Memory
  02.01.12 Knowledge
  02.01.13 Belief

III The social world
03 Society
  03.01 Society: the community
  03.01.01 Kinship/relationship
  03.01.02 Study of society
  03.01.03 Society in relation to customs/values/beliefs
  03.01.04 Social communication/relations
  03.01.05 Social attitudes
  03.01.06 Social classes/ranks
  03.01.07 Dissent/discred
  03.02 Inhabiting/dwelling
  03.02.01 Inhabiting/dwelling type of place
  03.02.02 Inhabiting/dwelling temporarily
  03.02.03 Providing with dwelling place
  03.02.04 Removing from dwelling place
  03.02.05 Furnishing with inhabitants
  03.02.06 Inhabitant/resident
  03.02.07 Inhabited place
  03.03 Armed hostility
  03.03.01 War
  03.03.02 Armed encounter
  03.03.03 Armed conflict

A general and abstract terms
B the body and the individual
C arts and crafts
D emotion
E money and commerce in industry
F food and farming
G government and public
H architecture, housing and the home
I numbers and measurement
J entertainment, sports and games
K life and living things
L movement, location, travel and transport
M social actions, states and processes
N education
O languages and communication
P psychology
Q science and technology
R the natural world
S social and cultural actions
T world and environment
U social actions, states and processes
V social actions, states and processes
W names and grammar
Historical Thesaurus of English

- Comprehensive analysis of English as found in the 2\textsuperscript{nd} edition of the OED
- 793,742 word forms arranged into 225,131 semantic categories
- The HT semantic categories are mapped to 4,028 thematic-level categories.
- three primary divisions are
  - I The External World
  - II The Mental World
  - III The Social World
- each category is given a nested reference code such as "01.02.08.02.02.06.01 n" for the category Whisky
Architecture of Annotation system

**Semantic Annotation System**

- **Input raw text**

- **Spelling train model**
- **USAS semantic lexicon resources**
- **Context-distance based algorithm**

- **VARD**
- **CLAWS**
- **USAS**
- **HT-based Sem. Tagger**

- **(SAMUELS Project rsc.)**
  - Historical Thesaurus;
  - Higher-level HT categories;
  - Linked HT categories;
  - Highly polysemous words;
  - Z-category words;

- **Annotated text**
<table>
<thead>
<tr>
<th>TOKEN</th>
<th>LEMMA</th>
<th>POSTAG</th>
<th>SEMTAG1</th>
<th>MWE</th>
<th>SEMTAG2</th>
<th>SEMTAG3</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>the</td>
<td>AT</td>
<td>Z5</td>
<td>0</td>
<td>04.03 [null];</td>
<td>04.03 [Grammatical Word];</td>
</tr>
<tr>
<td>cat</td>
<td>cat</td>
<td>NN1</td>
<td>L2 M3</td>
<td>0</td>
<td>04.10 [Unrecognised];</td>
<td>04.10 [Unrecognised];</td>
</tr>
<tr>
<td>sat</td>
<td>sit</td>
<td>VVD</td>
<td>M8 C1 P1 G1.1 G2.1 M6 A9+</td>
<td>0</td>
<td>[MWE] 01.02.08.01.22.08-13 [1.00000000] [Cook .burn/catch on bottom of cooking pot]; 01.05.08.09-06 [1.00000000] [Not moving .remain as opposed to go];</td>
<td>B24t07 [Cooking]; E08i [Absence/privation/cessation of movement];</td>
</tr>
<tr>
<td>on</td>
<td>on</td>
<td>II</td>
<td>Z5</td>
<td>0</td>
<td>[MWE] 01.02.08.01.22.08-13 [1.00000000] [Cook .burn/catch on bottom of cooking pot]; 01.05.08.09-06 [1.00000000] [Not moving .remain as opposed to go];</td>
<td>B24t07 [Cooking]; E08i [Absence/privation/cessation of movement];</td>
</tr>
<tr>
<td>the</td>
<td>the</td>
<td>AT</td>
<td>Z5</td>
<td>0</td>
<td>04.03 [null];</td>
<td>04.03 [Grammatical Word];</td>
</tr>
<tr>
<td>mat</td>
<td>mat</td>
<td>NN1</td>
<td>H5 O2</td>
<td>0</td>
<td>03.02.07.03.09.14-03 [0.93750000] [.mat]; 03.11.04.13.16.15-14 [0.93750000] [.mat]; 03.02.07.03.09.10.01-02 [0.94444444] [.table mat];</td>
<td>Q06f05m [Floor-covering]; Z08v11 [Bowls/bowling]; Q06f05i [Household linen];</td>
</tr>
<tr>
<td>.</td>
<td>PUNC</td>
<td>YSTP</td>
<td>PUNC</td>
<td>0</td>
<td>NULL</td>
<td>NULL [ ];</td>
</tr>
<tr>
<td>S_END</td>
<td>NULL</td>
<td>NULL</td>
<td>Z99</td>
<td>0</td>
<td>04.10 [Unrecognised];</td>
<td>04.10 [Unrecognised];</td>
</tr>
</tbody>
</table>
HTST current disambiguation methods (1)

• Disambiguate words and MWEs that have multiple HT categories
  – Filter by POS.
  – For each candidate category, extract all possible parent categories and collect headings (simple definition) of them, including current heading. Words in the headings form a feature set $\mathbf{HW}_i = \{h_1, h_2, ..., h_m\}$.
  – Collect up to five content words from each side of the key word/MWE. Together with the target word/mwe $w_t$, they form a context feature set $\mathbf{CW} = \{w_t, w_1, w_2, ..., w_n\}$.
  – Measure Jaccard Distance between $\mathbf{CW}$ and each $\mathbf{HW}_i$, and select the candidate categories (up to three) that have close distances to the context.
HTST current disambiguation methods (2)

- Time filtering
  - Filter word senses whose usage appear outside a given time window in the HT thesaurus.
  - Users can set upper and lower time boundaries (in years) to increase the relevance of the HT categories to the given time.
  - E.g. if a text was published in 1800, using the time filter, ignore the word senses which appear after that era.
  - Particularly useful for tagging historical data.
Further disambiguation methods

- Detecting linked HT categories in context to determine the core senses;
- Apply co-occurrence based statistical training model based on HT-OED sense mapping, OED example sentences (50.2M tokens) and sense definitions (14.5M tokens).
  - At word level: based on co-occurrence between HT category and context words
  - At semantic level: Based on co-occurrence between HT category and USAS tags.
- Core HT category detection based on density of polysemy;
- Core HT category detection based on OED sense ordering;
- Improve VARD with OED spelling variants data linked to headwords & dates.
Evaluation

• Ten texts were selected from different genres (e.g. spoken and written).
• Publication time spans from 1820 to 2014.
• Each text contains about 1,000 words.
• Evaluated for both HT sense codes and thematic sense codes.
• Examined the impact of the time filter.
• Evaluation criterion: If top three of the candidate tags suggested by the system contain the correct tag(s), it is considered to be correct annotation.
  – *In our evaluation, 81.18% and 12.74% of the correct tags were the first and second candidate tags respectively.*
Wmatrix main points

http://ucrel.lancs.ac.uk/wmatrix/

- Web-based (c.f. BNCweb, CQPweb)
- You can load your own (English) data
- Incorporates main methods in corpus linguistics toolbox
  - frequency lists, concordances, key words, collocations, n-grams
- Adds two levels of linguistic annotation (NLP or computational linguistics methods)
  - POS tagging, Semantic field tagging
- Novelty
  - key domain analysis, semantic collocations
Annotation (NLP): tag wizard

Corpus retrieval (CL): frequency lists, concordances, keyness, collocation, n-grams

MapReduce Pipeline

Decabillions Corpus Database
Cluster & cloud computing

- MapReduce (Hadoop) framework
- Hansard corpus processing
  - 2.2 billion words
  - 32.7GB of data including mark-up
  - 7.5 million files
  - 3 days to complete versus 98 days on one PC (HPC-USAS)
  - 6 days to complete on our hand-made cluster (HTST)
Big data/corpus challenges

• Our solutions
  – Variant Spelling methods
  – Improved semantic disambiguation techniques (Historical Thesaurus Semantic Tagger – HTST)
  – Use of big data methods e.g. cluster and cloud computing

• Future challenges
  – Visualisations / GIS
  – Improved use of statistical methods
  – Meeting the disciplinary needs of SSH researchers
Thanks for listening!

- p.rayson@lancaster.ac.uk
- @perayson

- http://www.gla.ac.uk/samuels/
- http://ucrel.lancs.ac.uk/vard/
- http://ucrel.lancs.ac.uk/usas/
- http://phlox.lancs.ac.uk/ucrel/semtagger/english