Computational Linguistics could Serve History of Science

John Nerbonne

Center for Language and Cognition
University of Groningen

Crossing Boundaries:
History of Science and Computational Linguistics
Bari, 28 April 2008

Outline

Acknowledgments

Ismail Fahmi, Ph.D. Candidate, Groningen
Automatic Term Recognition and Relation Identification for Medical Ontology Learning, ca. end 2008

Dr. Gosse Bouma, Fahmi's primary supervisor

Dr. Jörg Tiedemann, Jori Mur, Lonneke van der Plas, collaborators in information extraction in Groningen

Thesis

Computational Linguistics (CL) could serve the study of the history of science.

- Focus on information extraction, not on e.g. sentiment analysis
- CL now applied to extract essential ontologies, relations
    - not used to extract information on data selection, data preparation, analysis techniques, controls, qualifications, application ideas, mathematics, ... (lots of other scientific discourse)
- Techniques available, increasingly well understood
- Requires substantial amounts of text, preferable 10^6 words or more

Motivation for CL Work Practical

- Information Extraction (IE) seen as useful to many applications
- Indexing, summarizing, question-answering, providing “clippings”, reference works for students, practicioners, ...
- Most interest in extending useful techniques, not in reflecting on basic problems.

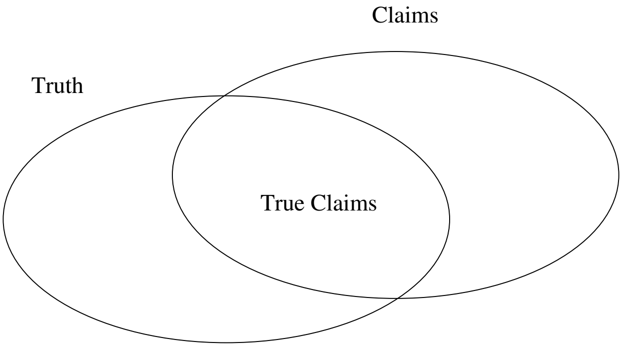
Goals of CL Work on Terminology

- Identify the terms used in a given domain, i.e. the words or phrases used to refer to the special concepts.

  renal dysplasia, kidney failure, renal infection, urinary tract/track infection, hydronephrosis, acute glomeruli-nehpritis, ...

- Identify the relations among the terms
  - *streptococcal pneumonia* is-a *bacterial pneumonia* is-a *pneumonia* is-a *lung disorder* is-a *cardio-pulmonary disorder* is-a ...
  - *high fever* is-a-symptom-of *streptococcal pneumonia*
  - *bacterial pneumonia* is-a-cause-of *glomeruli-nehpritis*

IR Evaluation



Precision = True-Claims/Claims, Recall = True-Claims/Truth

Evaluation, cont.

Problem: The “truth” may be unknown, at least in its entirety, i.e. all the concepts that might play a role.

- we rank the “claims” that are returned, i.e. the words and phrases we hypothesize to be terms
- at each rank, we measure the precision, i.e. how many of the words etc. up to that rank are genuine terms
- we report *average precision* for a certain number of ranks. These are often presented as scatterplots of precision in terms of “recall” rank.

Data Sources—Texts!

- Elseviers medical encyclopedia: a medical encyclopedia intended for general audience and containing 379K words. (courtesy of Spectrum, Ltd., online at http://www.kiesbeter.nl/medischeinformatie/
- Dutch edition of the *Merck Manual*, general-purpose medical handbook intended for professionals and containing 780K words (http://www.merckmanual.nl)

(Translated) Example

> **Acute necrotizing gingivitis**, *(***Vincent's stomatitis**, **acute necrotizing ulcerative gingivitis, (ANUG)***) is a painful, non-contagious infection of the gums that causes pain, fever and exahustion. It has also been known as* **trench mouth** *since the First World War, when many soldiers contracted the infection while on duty in the trenches ...*

Problem: identify the (bold-faced) terms, without misidentifying 'gums', 'soldiers', 'duty' or 'trenches' as terms

Preprocessing

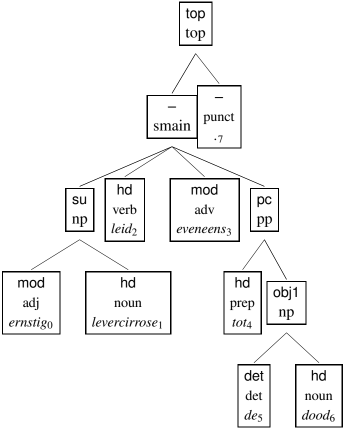
- For each document, removing formating codes, etc. to obtain "flat text". Exceptions: some structural (html) codes.
- Remove words outside normal prose such as titles and headings, keeping, however, incomplete sentences as those common at the beginning of definitional sections.
- (In Groningen) Analyse the sentences of the document syntactically using Van Noord's Alpino parser. (Others use part-of-speech tagging, etc.)
- "Treebank" of 60K parsed sentences

Simple Linguistic Analysis

- Tag the text automatically for parts of speech, i.e. syntactic categories of words.
- Search among POS-tag sequences using a regular expression (example simplified):
  ((Adj|N)+ | (((Adj|N)* (N Prep ...)?))) N

Finds sequences ending in a noun ‘N’ and preceded by one or more adjectives or nounds ‘(Adj|N)+’, or preceded zero or more adjectives or nouns, followed in turn by ... etc.

Alpino Linguistic Analysis



An analysis tree for the sentence *Ernstige levercirrose leidt eveneens tot de dood* (Severe liver cirrhosis also leads to death)

Appropriate Level of Analysis

Surprising result: simple, regular expression filtering outperforms sophisticated linguistic analysis (slightly).

...even though Alpino is one of the best parsing systems anywhere.

Identifying 2-Word *Units*

- Many terms consist of multiple words ‘acute gingivitis’, ‘Vincent’s stomatitis’, etc.
- Lots of research on recognizing multi-word units, two-word, three-word, ...
- Key idea: recognize when two words occur more than would be expected by chance: association strength reflects “unithoood”
- In a second step, use recognized terms to improve detection of compound terms.

Identifying *Units*

Aim: identify which combinations of words function as a unit, e.g. analysing which which words appear together more often than chance would predict.

	y	$\bar{y}$	
x	n_{11}	n_{12}	n_{1p}
$\bar{x}$	n_{21}	n_{22}	n_{2p}
	n_{p1}	n_{p2}	n_{pp}

Contingency table of frequency data for a word pair *xy*.
$p(xy) = n_{11}/n_{pp}$, $p(x) = n_{1p}/n_{pp}$ etc.

Measuring Association Strength

Raw frequency, χ^2, (pointwise) mutual information, log-likelihood, t-scores, ...

$$\mathrm{PMI}(x_i, y_j) = \log_2 \frac{p(x_i y_j)}{p(x_i)p(y_j)}$$

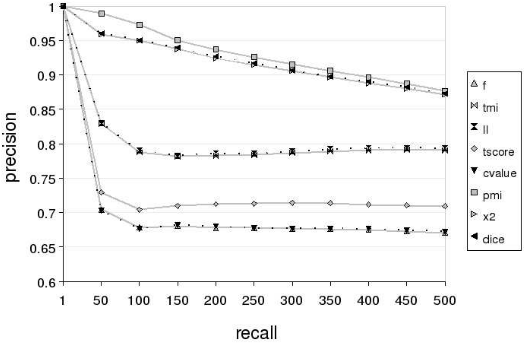
$$\mathrm{MI}(x, y) = \sum_{x,y} p(xy) \log_2 \frac{p(xy)}{p(x)p(y)}$$

Pointwise MI measures association strength between two concrete words (values in discrete distribution), MI sums over all values of variable.

Which statistical measure detects “unithood” best?

Evaluation Term Detection

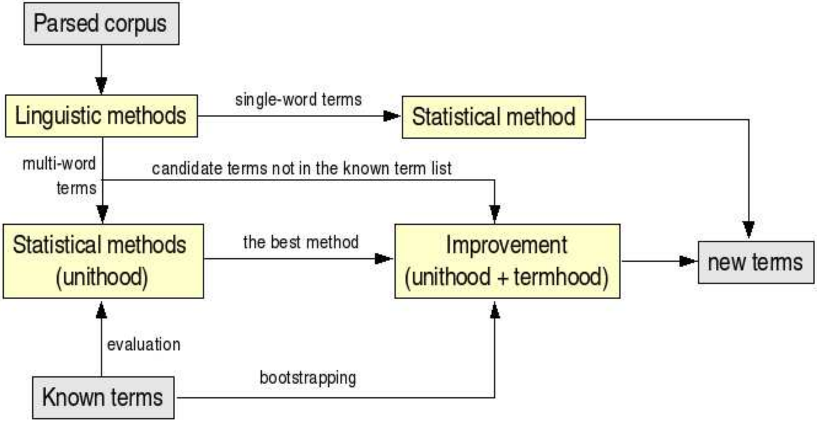
Comparing association strength measures, restricting attention to word pairs that appear at least eight times:



Other Components

- detect single-word terms via frequency in specialized corpora vs. general frequency
- use knowledge sources containing thousands of terms (Unified Medical Language System, UMLS)
- hypothesize that combinations using known terms will also be terms (“bootstrap”)
- detecting variations, synonyms, abbreviations

Identification Scheme



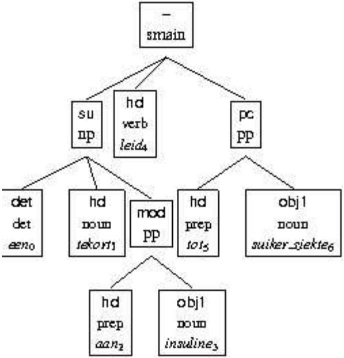
Relation Detection: Goals

- Given some identification of terms (see above)
- Identify the relations among the terms
  - *streptococcal pneumonia* **is-a** *bacterial pneumonia* is-a ...
  - Blood in the urine is a **symptom of** nephritis
  - Bacterial pneumonia often **occurs in** cases of glomeruli nephritis
  - Exercise and diet can **prevent** type-2 diabetes
  - ...

Relation Detection: Techniques

- we normally identify the relations of interest ahead of time (unlike the case of terms)
- exploit known terms (from term detection, or external knowledge source, or e.g. combined with detection of potential translation equivalence)
- seek clauses in which terms appear, note syntactic relations
- cluster clauses with similar configurations of terms (or learn classification)

Relation Detection



Evidence that ‘deficit of insulin’ stands in the ‘cause’ relation to ‘diabetes’.

Relation Detection Results

Relation type	Prec.	Recall
causes	0.83	0.73
occurs	0.81	0.54
has-symptom	0.58	0.62
prevents	0.80	0.40
treats	0.71	0.40
diagnoses	0.86	0.24

Hard problem in spite of large amounts of text

Conspectus

- CL is already extracting terminologies, which map to ontologies
- CL is making inroads to the problem of detecting relations among concepts, but it's still rough
- Large text resources are required
- One may not expect very high accuracy

Thesis

Computational Linguistics (CL) could serve the study of the history of science.

- Focus on information extraction, not on e.g. sentiment analysis
- CL now applied to extract essential ontologies, relations
    - not used to extract information on data selection, data preparation, analysis techniques, controls, qualifications, application ideas, mathematics, ... (lots of other scientific discourse)
- Techniques available, increasingly well understood
- Requires substantial amounts of text, preferable 10^6 words or more

Other Benefits

- tracking the introduction of terminology, ‘right’ in political discussion or ‘simulation’ in social sciences
- tracking the frequency distribution of terminology over time, ‘cognitive’ since the 1960’s, ‘genome’ since the 1990’s, ...
- tracking the social use of terminology—which sciences (and scientists) initiate terms that become popular?

Beyond Citation Analysis

Wendy Lenhert, Claire Cardie and Ellen Riloff "Analysing Research Papers Using Citation Sentences" *Cog.Sci.* 12, 1990, 511-518.

Classifies sentences in which citations occur (citing facts, applications, problems, criticisms, ...)

Political Science

Claire Cardie and John Wilkerson (eds.) “Text Annotation for Political Science Research”

special issue of *Journal of Information Technology & Politics*
CFP at journal web site

Case: SWHi

Junte Zhang, Ismail Fahmi, Henk Ellerman, & Gosse Bouma "Mapping Metadata for Semantic Web for History (SWHi)" *International Workshop on Collaborative Knowledge Management for Web Information Systems* 2007.

Populate ontology of historical figures using the *Early American Imprint Series I* in the format of library metadata

Little to no CL, but another example of how existing resources can be combined to extract information.