AUTOMATIC EXTRACTION OF FACTS FROM EPIDEMIOLOGICAL LITERATURE: A CASE STUDY OF THE MDC COHORT

JOHAN FRID, LUND UNIVERSITY HUMANITIES LABORATORY

JONAS BJÖRK, DIVISION OF OCCUPATIONAL AND ENVIRONMENTAL MEDICINE

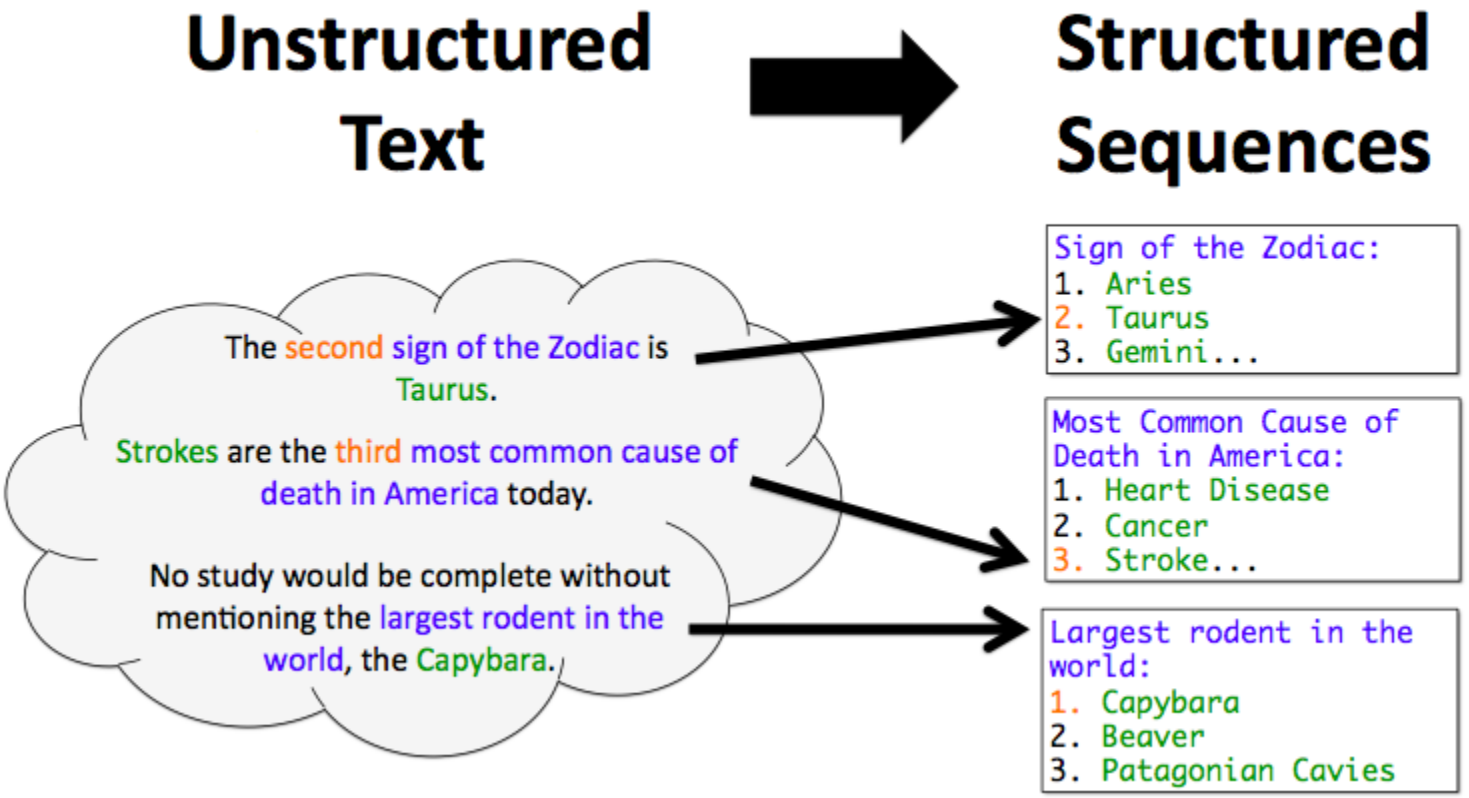
COMPUTE Winter Meeting, Lund 3 March 2017

Information Extraction

- Find and understand limited relevant parts of texts
- Gather information from many pieces of text
- Produce a structured representation of relevant information:
  - *relations* (in the database sense), a.k.a.,
  - a *knowledge base*
- Goals:
  - Organize information so that it is useful to people
  - Put information in a semantically precise form that allows further inferences to be made by computer algorithms
- E.g.,
  - Learn interactions drug-gene, exposure-outcome from medical research literature

Based on slide from Christopher Manning

Information extraction: facts



Information extraction: rhetorical categories

- IMRAD: Introduction, Method, Results and Discussion
- Sentence-based: italic represents introduction, underscore represents methods, bold represents results and italic-underscore represents discussion (Agarwal & Yu 2009)

PECAM-1 plays an important role in endothelial cell–cell and cell–matrix interactions, which are essential during vasculogenesis and/or angiogenesis (17,22). Here, we examined expression of PECAM-1 mRNA in vascular beds of various human tissues and compared it with expression of PECAM-1 in human endothelial and hematopoietic cells. ***A short exposure of the blot probed with GAPDH is shown, because poly(A)+ RNA from the cell lines gives a strong signal within several hours compared with the total RNA from human tissue. Therefore, total RNA from various tissues required a much longer exposure to reveal GAPDH mRNA.*** *Human tissue and cell lines expressed multiple RNA bands for PECAM-1, which may represent alternatively spliced PECAM-1 isoforms, the identity of which required further analysis.*

Problem formulation

- Our approach: hybrid, extension of IMRAD
  - Sentence based
- 13 epidemiological facts defined:
  - Aim, Exposures, Outcomes, Study design, Inclusion criteria, Actual size, Several cohorts, Subgroups, Follow-up period, Assessment of end-points, Statistical methods, Results, Conclusions
- Build a classifier that can tell us if:
  - a sentence explicitly *mention* or is in some way *related to* the [aim, exposure, outcome, …]?

Guidelines

- Document that describes and illustrates each category
- Aim
  - Sentences describing the overall objectives of the study. Usually relates to the content of an entire sentence rather than specific parts of it.
  - Example
    - This prospective study aimed to explore whether total and differential leukocyte counts areassociated with incidence of diabetes

Guidelines

- Exposures
  - Factors that the study individuals are exposed to and the study is designed to assess consequences of in terms of health outcomes or disease risk
  - Example
    - **Blood cadmium** levels were estimated from hematocrit and cadmium concentrations in erythrocytes
- Outcomes
  - Incident diseases or health conditions, that occur during the follow up of the study individuals
  - Example
    - Incidence of **diabetes** was studied during a mean follow-up of 14 years
- etc

Methods: Data set

- Search Plos One and Pubmed databases for papers containing all of the the words ’malmö’, ’diet’ and ’cancer’
- Download, apply Natural Language Processing (NLP)
- Total: 315 abstracts, 2984 sentences
- In this work: 92 abstracts, 1006 sentences
- Each sentence annotated indepentently by 2 researchers
  - Disagreements were discussed… and *solved*!

Methods: Annotation

- Turn the task into a binary decision task
  - Does this sentence explicitly *mention* or is it in any way *related to* the [aim, exposure, outcome, …]?
  - Mark sentence as *positive* (otherwise *negative*)
  - 13 categories = 13 decisions/sentence
- Similar to a popular NLP task - **sentiment analysis** - for which many methods exist

Interrator agreement

(number of positives per annotator and after syncing; n=1006. Agarwal & Yu: kappa=0.756)

Type	A1	A2	Cohen's Kappa	Sync A1,A2
Aim	93	63	0.778	91
Exposures	747	801	0.675	788
Outcomes	639	676	0.730	686
Study design	31	43	0.411	31
Inclusion criteria	26	27	0.554	31
Actual size	134	153	0.866	154
Several cohorts	53	67	0.628	62
Subgroups	138	115	0.499	142
Follow-up period	57	63	0.823	68
Assessment of end-points	16	42	0.471	16
Statistical methods	73	66	0.822	81
Results	327	317	0.813	323
Conclusions	159	106	0.607	164

Methods: NLP

- CoreNLP (Stanford)
  - http://stanfordnlp.github.io/CoreNLP/
- Sentence splitting
  - Divide each abstract into sentences
- Lemmatization
  - Standardise singular and plural forms, different verb tenses
- Named Entity Recognition (NER)
  - Standardise numeric strings, genes

Methods: Bag-of-n-grams

(sentence ”vectorization”)

```
corpus = [
'This is the first document.',
'This is the second second document.',
'And the third one.',
'Is this the first document?‘
]
```

```
ngrams2 = ['and', 'and the',
'document', 'first', 'first
document', 'is', 'is the', 'is
this', 'one', 'second',
'second document', 'second
second', 'the', 'the first',
'the second', 'the third',
'third', 'third one', 'this',
'this is', 'this the']
```

analyse = ['This is a text document to analyze.']

and	and the	document	first	first document	is	is the
0	0	1	0	0	1	0
is this	one	second	second document	second second	the	the first
0	0	0	0	0	0	0
the second	the third	third	third one	this	this is	this the
0	0	0	0	1	1	0

Methods: Machine learning

- NBSVM (Wang & Manning 2012; sentiment analysis)
  - Compute a log-ratio vector between the n-gram counts extracted from positive examples and the n-gram counts extracted from negative examples
  - Input to the classifier is the log ratio vector multiplied by the binary pattern created from the n-grams in the example
  - Classifier can be SVM or logistic regression, we use linear SVM
    - Liblinear (https://www.csie.ntu.edu.tw/~cjlin/liblinear/)

Methods: Bag-of-n-grams

```
corpuspos = [
‘I am positive.‘
]
corpusneg = [
‘I am negative.‘
]
```

```
ngram, log-ratio
’i am positive', 0.69
’i', 0
’positive', 0.69
’am', 0
’negative', -0.69
’am positive', 0.69
’am negative', -0.69
'i am negative', -0.69
’i am’, 0
```

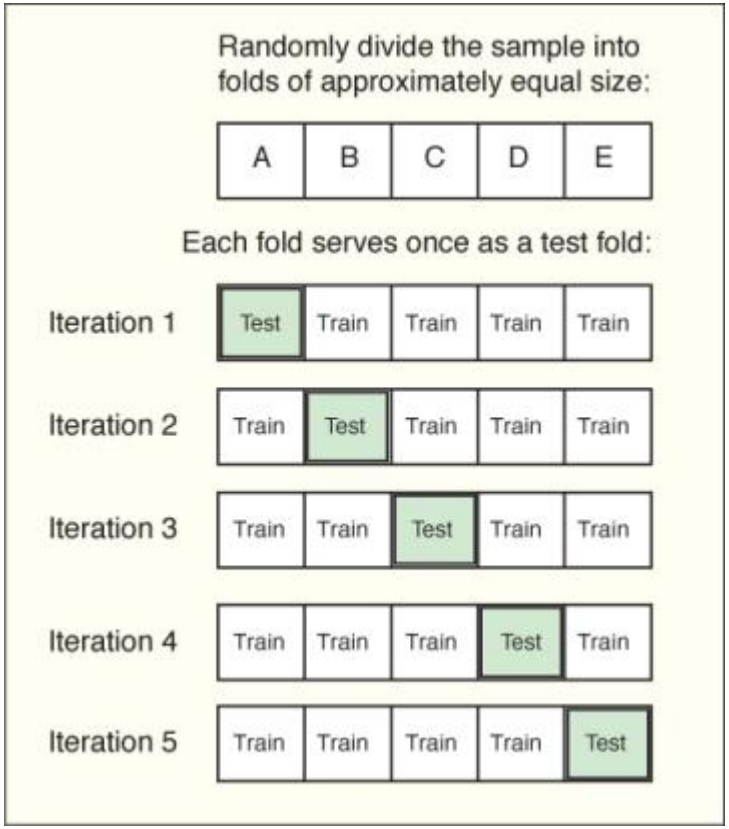
analyse = [‘I am negative too.']

i am positive	i	positive	am	negative	am positive	am negative
0.69*0	0*1	0.69*0	0*1	-0.69*1	0.69*0	-0.69*1

i am negative	i am	value from model***value from analysis sentence**
-0.69*1	0*1	’too’? It’s not in the model!

Methods: Evaluation

- 10-fold cross-validation
  - Train: ~900 sentences
  - Test: ~100 sentences



Methods: Evaluation

- Binary decision
  - yes or no
- Confusion matrix

accuracy (ACC)

$$ACC = (TP + TN)/(TP + FP + FN + TN)$$

F1 score

is the harmonic mean of precision and sensitivity

$$F1 = 2TP/(2TP + FP + FN)$$

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = Σ True Positive / Σ Test Outcome Positive
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = Σ True Negative / Σ Test Outcome Negative
		Sensitivity = Σ True Positive / Σ Condition Positive	**Specificity** = Σ True Negative / Σ Condition Negative	

Results

(Agarwai & Yu: F1=0.9155; Wang & Manning: Acc=0.89-0.93)

Type	Sensitivity	Specificity	F1	Accuracy	Accuracy (0R)	Acc/Acc(0R) in %
Aim	0.52	0.99	0.68	0.95	0.92	3.83
Exposures	0.98	0.41	0.92	0.86	0.79	9.14
Outcomes	0.96	0.59	0.89	0.84	0.69	22.93
Study design	0.06	1.00	0.12	0.97	0.98	-0.51
Inclusion criteria	0.07	1.00	0.13	0.97	0.98	-0.51
Actual size	0.78	0.99	0.87	0.96	0.85	12.68
Several cohorts	0.28	1.00	0.44	0.95	0.94	0.64
Subgroups	0.16	0.98	0.26	0.87	0.86	0.31
Follow-up period	0.58	0.99	0.73	0.97	0.94	3.06
Assessment of end-points	0.06	0.99	0.11	0.98	0.99	-1.01
Statistical methods	0.59	1.00	0.74	0.96	0.93	3.78
Results	0.76	0.96	0.83	0.89	0.68	30.80
Conclusions	0.19	0.99	0.31	0.85	0.84	1.35

Application to the MDC corpus

- Numeric results are nice - but does it *work*?
- Search the rest 2664 sentences
  - Demo: https://goo.gl/QqV9FV
- Exposures and Outcomes are really easy to find (high sensitivity), but also false negatives (lower specificity)
- What about the others?

Application to the MDC corpus

Actual size	Follow-up
HB-EGF , EGF and platelet-derived growth factor were measured in plasma from 202 patients undergoing carotid endarterectomy and in 384 incident CE cases and 409 matched controls recruited from the Malmö Diet and Cancer cohort . To replicate the results , we also assessed another large independent cohort cross-sectionally , the Malmö Diet and Cancer Study -LRB- MDC , n = 26 777 -RRB- . In addition , the prognostic value of ezrin expression is validated in primary tumours , and the longitudinal expression of ezrin examined in a subset of primary and recurrent tumours -LRB- n = 28 -RRB- . A random sample of participants -LRB- age , 45-68 years -RRB- in the population-based Malmö Diet and Cancer Study underwent B-mode ultrasound with measurements of IMT and the presence of plaque in the common carotid artery -LRB- n = 5079 -RRB- .	Incidence of ischemic stroke was monitored over a mean follow-up of 14.9 ± 3.0 years . During a mean follow-up of 14 y , 2648 CVD cases were identified . All patients were followed until 31 December 2010 using the Swedish Cause of Death Registry . The incidence of first cardiovascular events -LRB- myocardial infarction and stroke -RRB- and cause-specific mortality were monitored over a mean follow-up period of 13.2 years . Incident cases of HF and AF were identified from the Swedish hospital discharge register during a median follow-up of 16.3 years .

Application to the MDC corpus

Statistical methods	Assessment of end-points
Cox proportional hazards regression estimated hazard ratios -LRB- HRs -RRB- and 95 % confidence intervals -LRB- CIs -RRB- of breast cancer associated with fibre and 11 plant food groups . The likelihood of being a high consumer of HCAs was estimated by logistic regression analysis . Kaplan Meier analysis and Cox proportional hazards modelling were used to assess the relationship between RBM3 and recurrence free survival -LRB- RFS -RRB- and overall survival -LRB- OS -RRB- . Analyses were performed in quartiles of baseline age , and linear trends in effect size across age groups were estimated in logistic regression models . Associations with cancer-specific survival -LRB- CSS -RRB- were explored by Cox proportional hazards regression , unadjusted and adjusted for age , TNM stage , differentiation grade , vascular invasion and microsatellite instability -LRB- MSI -RRB- status .	Cancer incidence was assessed in both the SOS and the MDC cohorts through national and local registers . Individuals with a first AF diagnosis were identified by record linkage with national registers . Cancer cases and clinical characteristics were ascertained through national and regional registry data . Cases and clinical characteristics were ascertained via national and regional registry data . Cancer incidence and cause of death were retrieved using record linkage with national registries .

Application to the MDC corpus

Results	Conclusions
High-fibre bread was significantly associated with a decreased breast cancer incidence -LRB- HR , 0.75 ; 95 % CI , 0.57-0 .98 , for highest compared with lowest quintile -RRB- .	The same enzyme is implicated in ω-hydroxylation of very long and medium chain fatty acids in the liver suggesting its possible influence on gluco-metabolic components of MetS .
Membranous but not cytoplasmic ezrin was significantly reduced in recurrent compared to primary tumours -LRB- p < 0.001 -RRB- .	This association appears to be stronger in women with symptomatic cancers but breast density could not explain differences in survival according to detection mode .
The OR for high compared with moderate plasma alkylresorcinol metabolites was 1.41 -LRB- 95 % CI , 1.10-1 .80 -RRB- for prostate cancer .	Results from this large study with high-validity dietary data suggest that a high intake of refined carbohydrates may be associated with increased risk of prostate cancer .
After full multivariate adjustment , MR-ANP was significantly inversely associated with incident diabetes -LRB- OR = 0.85 ; 95 % CI = 0.73-0 .99 ; P = 0.034 -RRB- but not N-BNP -LRB- OR = 0.92 ; 95 % CI = 0.80-1 .06 ; P = 0.262 -RRB- .	Low plasma levels of MR-ANP predict development of future diabetes and glucose progression over time , suggesting a causal role of ANP deficiency in diabetes development .
	The potential utility of RBM3 in treatment stratification of patients with melanoma should be pursued in future studies .

N-gram evaluation – the good

(most positive log ratios)

Aim	was_*_to to_*_investigate aim
Exposures	association_*_between dietary cadmium
Outcomes	risk_*_of incidence breast
Study design	case-control_*_study nested_*_case-control_*_study a_*_nested_*_case-control
Inclusion criteria	without_*_history_*_of without_*_history NUMBER_*_individuals_*_with
Actual size	NUMBER_*_controls and_*_NUMBER_*_controls NUMBER_*_subjects
Several cohorts	epic prospective_*_investigation_*_into prospective_*_investigation
Subgroups	pinteraction_*_=_*_NUMBER pinteraction_*_= pinteraction
Follow-up period	during_*_a follow-up_*_of follow-up_*_of_*_NUMBER
Assessment of end-points	registers national were_*_identified_*_from
Statistical methods	regression cox proportional
Results	NUMBER_*_%_*_ci %_*_ci p_*_=_*_NUMBER
Conclusions	our_*_results provide suggests_*_that

N-gram evaluation – the bad

(most negative log ratios)

Aim	NUMBER_*_, p ,_*_NUMBER
Exposures	diagnosed_*_with were_*_diagnosed diagnosed
Outcomes	history_*_method diet_*_history_*_method diet_*_history
Study design	--_*_NUMBER NUMBER_*_% %
Inclusion criteria	risk_*_of ci NUMBER_*_-lrb-
Actual size	confidence increased mortality
Several cohorts	risk_*_of hr increased
Subgroups	to_*_NUMBER used plasma
Follow-up period	intake factors is
Assessment of end-points	risk_*_of association to
Statistical methods	NUMBER_*_; increased associated_*_with
Results	prospective may using
Conclusions	NUMBER_*_-lrb- -rrb-_*_, ,_*_NUMBER

N-gram evaluation – the meh

(most close to 0; selected examples)

Aim	rate relationship_*_between identify
Exposures	mean cancer_*_study
Outcomes	diet malmö baseline
Study design	NUMBER_*_controls
Inclusion criteria	controls NUMBER_*_cases
Actual size	follow-up diet cohort
Several cohorts	into data_*_from swg
Subgroups	men obesity fiber
Follow-up period	time mean of_*_number
Assessment of end-points	were_*_identified birth
Statistical methods	,_*_sex used hazard
Results	with associated not was
Conclusions	that these increased_*_risk_*_of

Discussion: Limitations

- Sentences currently treated in *isolation*; evidence for p and n may come from surrounding sentences
- Categories treated in isolation; may be relationships
- Abstracts; does it work for full-text?
- Other categories: interaction, effect size
- Unbalanced categories

Discussion: Future work

- Extend to actual fact extraction
- NLP: noun phrases, more NER
- Sentence vectorization: Word vectors
  - really ’hot topic’ in NLP
- Machine learning: RNN, CNN
- Evaluation: area under ROC curve

Test sentences

- 1 The inflammatory mediator procalcitonin -LRB- PCT -RRB- has previously been associated with prognosis in myocardial infarction , cancer and sepsis patients .
- 2 The importance of PCT in the general population is currently unknown .
- 3 Our aim was to assess the relationship between plasma PCT and the risk of all-cause and cause-specific mortality in apparently healthy individuals with no previous history of cardiovascular disease or cancer .
- 4 We performed a prospective , population-based study on 3,322 individuals recruited from the Malmö Diet and Cancer cohort , with a median follow-up time of 16.2 years .
- 5 Plasma PCT , high-sensitivity C-reactive protein -LRB- hsCRP -RRB- , low-density lipoprotein -LRB- LDL -RRB- , high-density lipoprotein -LRB- HDL -RRB- , triglycerides and cystatin C were measured at baseline and a thorough risk factor assessment was performed for all subjects .
- 6 The primary end-points of the study were all-cause mortality , cancer mortality and cardiovascular mortality .
- 7 Men had higher PCT levels compared to women .
- 8 In Cox proportional hazard models adjusted for age , sex , hypertension , diabetes , plasma lipids , renal function , body mass index and smoking , baseline PCT was associated with all-cause mortality and cancer mortality in men .
- 9 The hazard ratio -LRB- HR -RRB- for men with PCT levels within the highest compared with the lowest quartile was 1.52 -LRB- 95 % confidence interval -LRB- CI -RRB- 1.07 to 2.16 ; P = 0.024 -RRB- for all-cause mortality and 2.37 -LRB- 95 % CI 1.36 to 4.14 ; P = 0.006 -RRB- for cancer mortality .
- 10 Additionally , men with increased plasma PCT were found to be at a higher risk to develop colon cancer -LRB- HR per 1 SD increase = 1.49 -LRB- 95 % CI 1.13 to 1.95 -RRB- ; P = 0.005 -RRB- .
- 11 In multivariate Cox regression analyses with mutual adjustments for PCT and hsCRP , PCT was independently associated with cancer death -LRB- HR per 1 SD increase = 1.28 -LRB- 95 % CI 1.10 to 1.49 -RRB- ; P = 0.001 -RRB- and hsCRP with cardiovascular death -LRB- HR per 1 SD increase = 1.42 -LRB- 95 % CI 1.11 to 1.83 -RRB- ; P = 0.006 -RRB- in men .
- 12 We found no significant correlations between baseline PCT or hsCRP and incident cancer or cardiovascular death in women .
- 13 We disclose for the first time important independent associations between PCT and the risk for all-cause and cancer mortality in apparently healthy men .
- 14 Our findings warrant further investigation into the mechanisms underlying the relationship between PCT and cancer .