Distilling Conceptual Connections from MeSH Co-Occurrences

Padmini Srinivasan, Dimitar Hristovski

presented @ MEDINFO 2004

15 February 2005

INLS 279 Bioinformatics Seminar

Patrick Herron

SILS, UNC Chapel Hill

Goals

- Analyze MeSH heading/subheading pair co-occurrences, *e.g.*,

  (diabetes / drug therapy) with (chemical / therapeutic use)
  heading/concept *subheading* *heading/concept* *subheading*

- *Interestingness*: Select semantically meaningful ones (via chi-sq) that are relatively *domain independent*
- Develop a “reasonable representation” of each pair: “a weighted vector […] [emphasizing] verb based functional aspects of the underlying semantics”
- **The ultimate goal: such pairs may aid in generating connections across disciplines across all of MEDLINE**

Reducing the problem space: using the SN

- In order to reduce scale of the problem space…
  20,742 concepts * 82 headings=1.7 mil heading/subheading pairs,
  $1.7\text{mil}^2/2$ (number of possible heading subheading pairs) = 1.5 trillion
- and in order to be as domain-independent as possible…
- Represent concepts by their semantic types in the SN
- 20,742 -> 134 (semantic type/subheading)
- Problem space down to
  $(134*82)^2/2$ = 60 mil (two orders of magnitude smaller)
- Only 1 mil of that 60 mil is meaningful

Method

- Usual approaches vs. Srinivasan et al
- Extraction of [$(st/sh)_i$, $(st/sh)_j$] pairs
- Selection of background dataset
- Analysis of a single pair

Usual approaches

- First hand-pick relevant verbs & then extract their arguments
- Set of 1 mil co-occurrences too big for manual approach
- We want to extract interesting verbs based on MeSH co-occurrence

Srinivasan *et al* approach

- Two step approach:
  1. automatically identify/extract key verbs associated with a [$(st/sh)_i$, $(st/sh)_j$] pair
  2. Use these verbs to extract highly related Ns and NPs
- This paper establishes one method for step 1; work on step 2 is for a later date
- We want a weighted verb vector for MeSH co-occurrences

Extraction of [$(st/sh)_i$, $(st/sh)_j$] pairs

- Corpus: MEDLINE to 2001 into rows with MEDLINE record id (MRI), Head/subhead
- Transformed into MRI, Sem Type/subhead
- Then [$(st/sh)_i$, $(st/sh)_j$] pairings picked & frequencies noted
- 30% 1x, 97% < 500 over 11 million records
- Pairings further culled by two more criteria: 1. freq > 500) down to 31,000 pairs & 2. (observed co-occurrence >= 1.25*expected co-occurrence) lowered total to 22,000 pairs
- 250 randomly selected; documents were reliably retrieved for 228 of those

Co-occurrence calculations

- Actual co-occurrence of pair A,B

  (# docs w/ A * # docs w/B) / (total number of documents)^2

- Expected co-occurrence of pair A,B

  (# docs w/ A * # docs w/B) / total number of documents

- (Expected – observed)/expected * 100 > 0.25

Selection of background dataset

- 100,000 MEDLINE records randomly selected
- Title + abstract POS-tagged
- Verbs extracted & used as vector to represent doc; verbs transformed to infinitives
- IDF for each verb as $\log_2(100{,}000/df)$
- BV: background vector: set of (verb, IDF) pairs for each record – our doc vector D???

Analysis of a single pair

- Identify all docs in which pair appears
- 2/3^{rds} of docs placed in training set
- Other 1/3rd plus random 1/3rd from MEDLINE in test set
- Create verb profile for pair
- Test verb profiles

Creating verb profile

- Docs POS-tagged & Vs extracted
- Rules for inclusion: V must occur in at least 5 docs; V frequency in training set must be significantly different from its freq in BV – using Pearson's χ^2 test (null hypothesis: difference between expected and observed frequencies is random)
- Formation of the profile vector, to which weights are assigned just like with the BV, BUT..

Formation of the profile vector

- Four different profile vectors are formed:
  - V,AugTF
  - V,AugTF*IDF
  - V,TF*IDF
  - V,IDF
- Empirical question as to how each performs

Testing profiles

- Test each of the 4 PVs against the doc vectors D of the random set and that 1/3rd of documents in which pairs occur
- Similarity of a background vector to a PV as dot product of two vectors (D, PV)
- Mean similarity (D, PV) of random set & mean similarity of topic set calculated
- Variance/information/interestingness: significant difference in similarities for a pair indicates interestingness

Results: Verb profiles

- Example of top five verbs from the AugTFIDF PV

Pair	Verbs
[(disease or syndrome, drug) & (lipid, adverse effect)]	Withdraw, warrant, undertake, treat, tolerate

- Authors claim these verb profiles will provide useful constraints for extracting pair-associated nouns

Results: test set

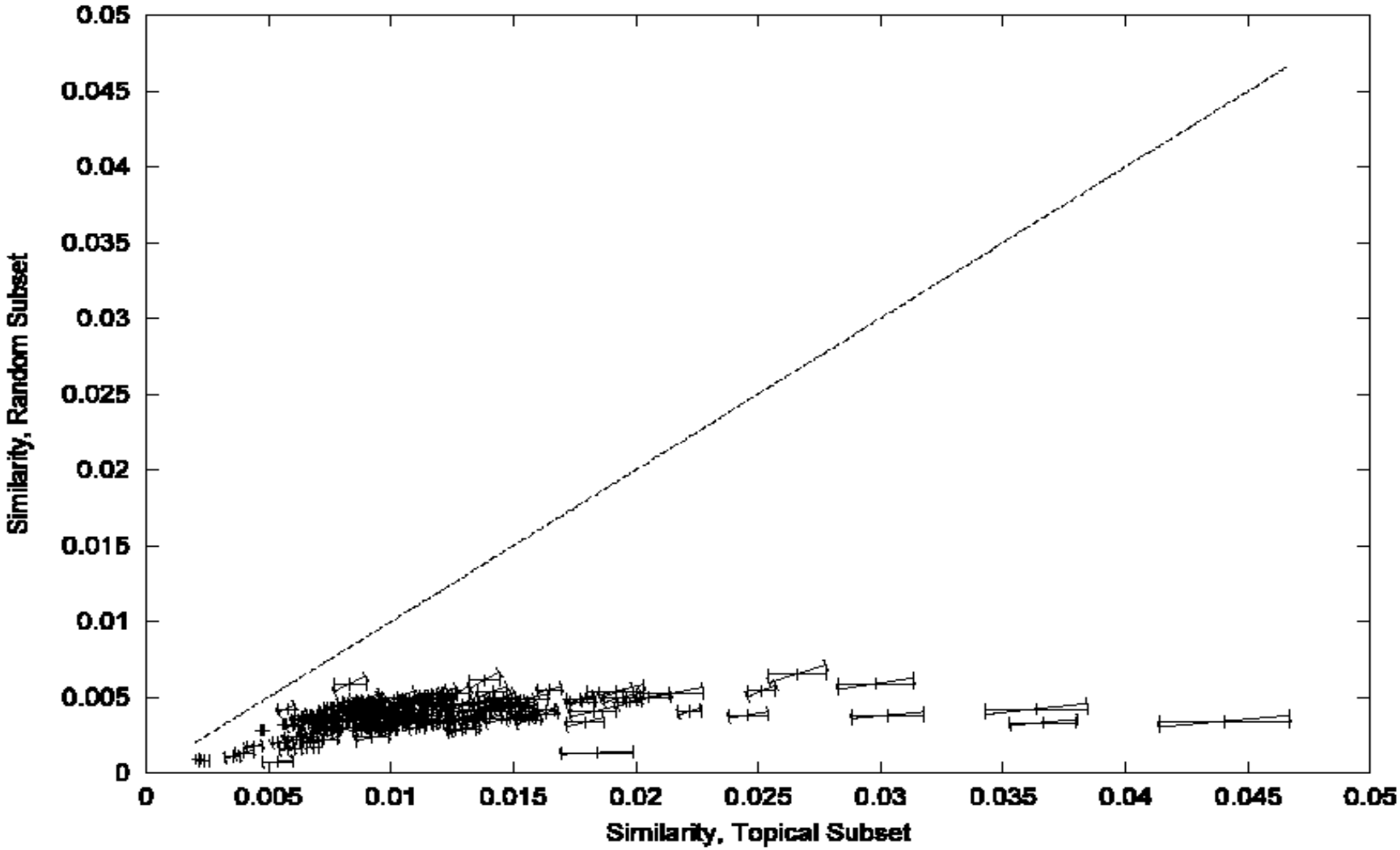

Figure 1 - Augmented TF.

Results: evaluation

- Each bar represents one of the 228 pairs plus standard error
- If a line touches the diagonal then the similarity is possibly random
- All pairs show significant similarity
- “in the right direction”

Conclusions

- Verbs are important
- Verb profiles from docs with MeSH co-occurrence pairs are different from docs not covered by pair: verb profiles can be used to characterize other docs w/ co-occurrenc

Questions

- How did Srinivasan et al decide only 1 mil of the 60 mil possible [(st/sh), (st/sh)] pairings are meaningful? Chi sq test with null hypothesis that pairings are random?
- What is the meaning of those similarity values? *How* significant???
- Once we have noun-verb representations of MeSH co-occurrences, what does that get us?
- What's truly interesting about MeSH co-occurrence? Connecting otherwise disparate pieces of the literature