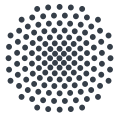
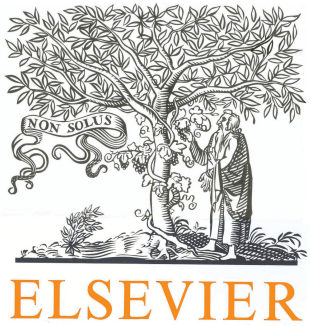


University of Stuttgart
Institute for
Natural Language Processing



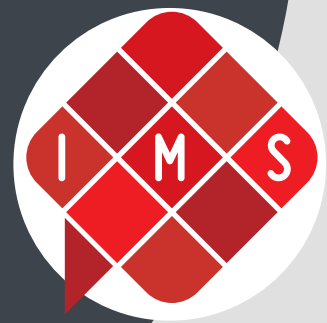
On the Semantic Similarity of Disease Mentions in Medline and Twitter

NLDB, Paris, France, 13th of June 2018

Camilo Thorne[1,2] **Roman Klinger**[2]

[1]Elsevier, [2]Universität Stuttgart

Slides are on
`http://www.romanklinger.de/talks`



Outline

1. Motivation
2. Experiments
   - Datasets
   - Quantitative Analysis
3. Results
   - Quantitative
   - Qualitative
   - Conclusion

Outline

1. Motivation
2. Experiments
   - Datasets
   - Quantitative Analysis
3. Results
   - Quantitative
   - Qualitative
   - Conclusion

Disease Names on Twitter

D009369 (Neoplasms), synonym: “cancer”

“Diabetes and Obesity Linked to Higher **Cancer** Risk: 4 Foods That Reduce This Risk”

D001321 (Autistic Disorder), synonym: “autism”

“UK Study: Brains Of Children With **Autism** Are Loaded With Aluminium”

D003865 (Depressive Disorder, Major), synonym: “SAD”

“Trot racing.. **sad** thing is I did not know they were going to do it. ... it’s only a rort....if...”

Goal (1)

Long Term Goal

- (Fake) Health news detection (on Twitter/Social Media)
- Pharmacovigilance (on Twitter/Social Media)

Prerequisites

Ability to recognize phrases...

- ...that refer to a given disease: a concept
- ...with one out of several synonyms

Goal (2)

Challenge

- How to query and annotate Tweets?
- Can we use tools existing for scientific literature?

Research Questions

- Can synonyms known from the scientific literature be used?
- Can distributional semantics uncover challenging cases?

Outline

1. Motivation
2. Experiments
   - Datasets
   - Quantitative Analysis
3. Results
   - Quantitative
   - Qualitative
   - Conclusion

Corpus Preparation and Collection

Medline

- abstracts between 01–2007 to 12–2017
- diseases detected and normalized to MeSH and OMIM using DNorm(⇐ important!)

Twitter

- queried tweets between 12–2017 and 03–2018
- query terms: 20 synonyms of top 100 most frequent concepts in Medline corpus

Statistics

Corpus	#Tokens	Units	Concepts	Synonyms
Medline	1,037,482,692	5,374,700	8,386	2,190,522
Twitter	145,793,358	7,193,077	4,908	201,712

Quantitative: How specific are diseases?

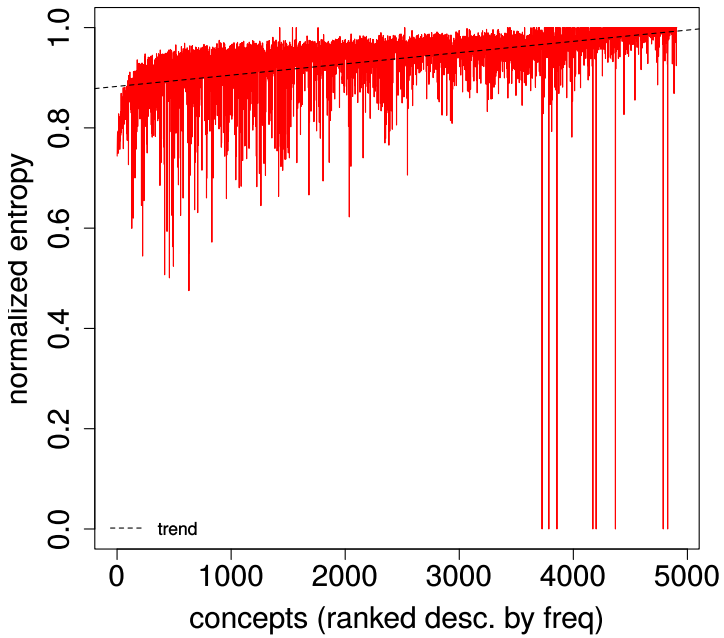
- Represent concepts/synonyms d as **distributional vectors** $\vec{D}$
  $\Rightarrow$ count window of +/– 5 words
- Operationalize semantic **specificity** as **normalized entropy**

$$H(d) = -\frac{1}{n}\sum_{w_i}^{n} P(w_i) \cdot \log_2 P(w_i)$$

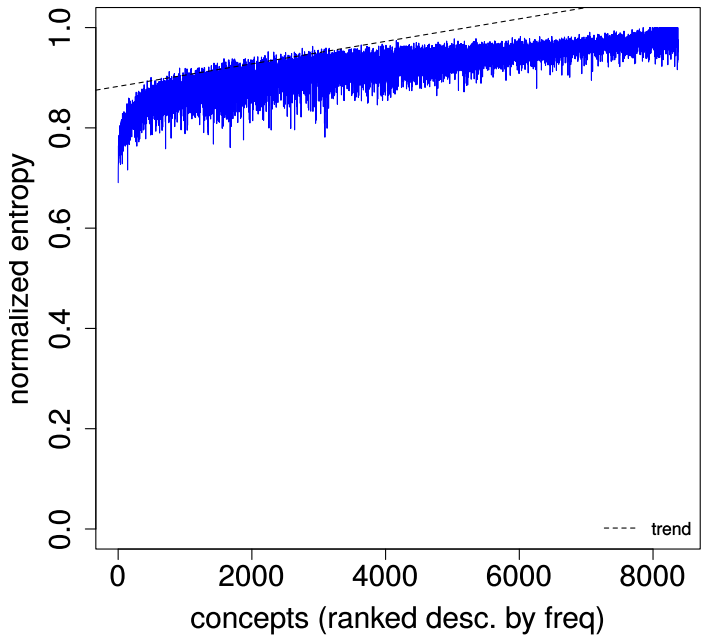
$\Rightarrow$ Measure is independent of context sizes or frequencies

$\Rightarrow$ The higher $H(d)$ the more **ambiguous** d

Quantitative: How specific are diseases?



Twitter



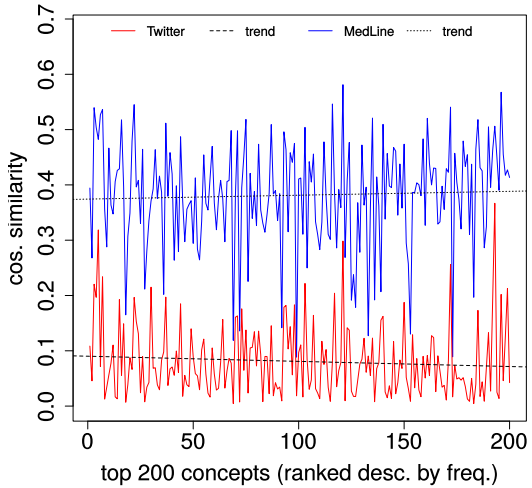
Medline

- Frequent concepts have lower ambiguity (significant on both Medline and Twitter)
- Ambiguity higher in Twitter than in Medline

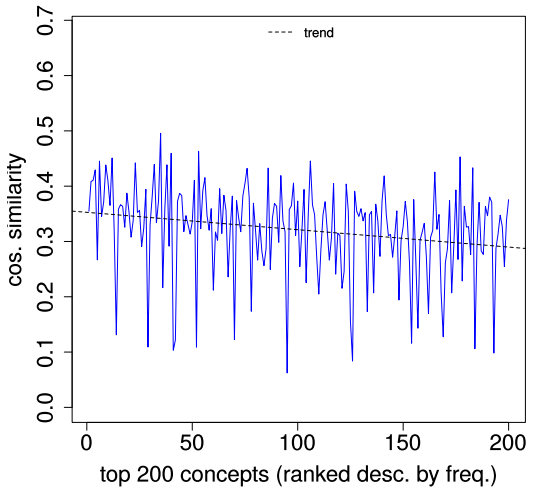
Quantitative: How similar are diseases (Twitter/Medline)?

- Use joint distributional model across both corpora
- Similarity of two concepts/synonyms: Cosine similiarity
- Analyse Similarity between top 200 concepts
- Analyze spread of the top 20 synonyms in each concept

Similarity – Results



Intra-concept Synonym Similarity



Concept Similarity

- Intra-concept similarity stable (avg.) w.r.t. frequency
- Intra-concept similarity higher on Medline than in Twitter
- Twitter-Medline similarity decreases with frequency (sign.)
- Twitter synonyms are less similar to each other than on Medline
- Concept similarity between M/T increases with frequency

Qualitative: Most/least similar concepts

Sim	Canonical name
0.496	Hepatitis C
0.463	Glioma
0.459	Diabetes Mellitus
0.453	Chronic Hepatitis
0.451	Hypoxia
0.446	Coronary Disease
0.445	HIV Infections
...	...

Sim	Canonical name
...	...
0.170	T Cell Leukemia
0.155	Cerebral Palsy
0.122	Van der Woude syndrome
0.116	Congenital bilateral aplasia of vas deferens
0.109	Sialic Acid Storage Disease
0.109	BMD

- **common** diseases have similar cross-corpus meaning
- **rare** diseases have dissimilar cross-corpus meaning

Qualitative: Synonym Level

Top 3 and bottom 3 synonyms of **Multiple Myeloma** in Medline and Twitter:

	Synonym	Entropy	Freq.
Medline	myeloma	0.807	10,706
	multiple myeloma	0.830	4,559
	AL	0.867	3,684
	extramedullary myeloma	0.936	15
	myeloma tumors	0.956	16
	lymphoma	0.944	16
Twitter	myeloma	0.868	1,787
	multiple myeloma	0.832	525
	Myeloma	0.911	389
	myelomas	1.000	5
	myeloma diagnosis	0.989	4
	Gamida	0.914	4

Top 3 and bottom 3 synonyms of **Angelman Syndrome** in Medline and Twitter:

	Synonym	Entropy	Freq.
Medline	AS	0.813	24,585
	AS-OCT	0.872	615
	Angelman syndrome	0.856	422
	AS-PC	0.948	8
	AS-AIH	0.932	8
	AS infection	0.931	9
Twitter	AS	0.927	598
	happiness	0.901	483
	Happiness	0.850	135
	Militer AS	0.976	3
	AS A CHILD	0.968	3
	Angelman Syndrome	0.947	3

False positive hits are a problem on Twitter.

Conclusions

- First distributional analysis of disease mentions on Twitter and Medline
- Concepts are dissimilar if uncommon
- Existing tools might be usable for popular diseases, but not across the full array of diseases
- Careful selection of synonyms used for Twitter analysis necessary, especially for less common diseases

Thank you! Questions?