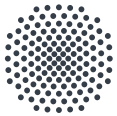
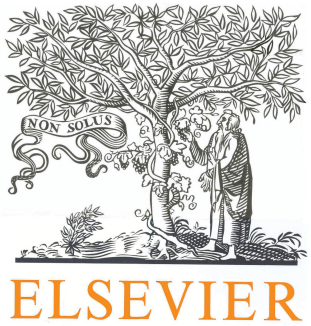


University of Stuttgart
Institute for
Natural Language Processing



On the Semantic Similarity of Disease Mentions in MEDLINE and Twitter

NLDB, Paris, France, 13th of June 2018

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

[1]Content and Innovation, Elsevier

[2]Institut für Maschinelle Sprachverarbeitung,
Universität Stuttgart

Outline

1. Motivation
2. Experiments
   - Datasets
   - Quantitative Analysis
   - Qualitative Analysis
3. References

Outline

1. Motivation
2. Experiments
   - Datasets
   - Quantitative Analysis
   - Qualitative Analysis
3. References

Disease Names on Twitter

Disease Names on Twitter

D009369 (Neoplasms), synonym: “cancer”

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

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 `https://t.co/3mht1SPlis`”

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 `https://t.co/3mht1SPlis`”

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... `https://t.co/sFnSLA3vJC`”

Goal (1)

Long Term Goal

Goal (1)

Long Term Goal

- Fake health news detection (on Twitter/Social Media)

Goal (1)

Long Term Goal

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

Goal (1)

Long Term Goal

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

Prerequisites

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 canoical disease a concept

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 canoical disease a concept
- ...with one out of several synonyms

Goal (2)

Challenge

Goal (2)

Challenge

- How to query and annotate Tweets?

Goal (2)

Challenge

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

Goal (2)

Challenge

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

Research Questions

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?

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
   - Qualitative Analysis
3. References

Corpus Preparation and Collection

MEDLINE

- abstracts between 01-2007 to 12-2017
- diseases detected and normalized to MeSH and OMIM UIDs using DNorm

Corpus Preparation and Collection

MEDLINE
- abstracts between 01-2007 to 12-2017
- diseases detected and normalized to MeSH and OMIM UIDs using DNorm

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

Corpus Preparation and Collection

MEDLINE
- abstracts between 01-2007 to 12-2017
- diseases detected and normalized to MeSH and OMIM UIDs using DNorm

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?

Quantitative: How specific are diseases?

- Represent concepts/synonyms d as **distributional vectors** $\vec{D}$
  $\Rightarrow$ count sliding window of +/- 5 words

Quantitative: How specific are diseases?

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

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

Quantitative: How specific are diseases?

- Represent concepts/synonyms d as **distributional vectors** $\vec{D}$
  $\Rightarrow$ count sliding 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 from context sizes or frequencies

Quantitative: How specific are diseases?

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

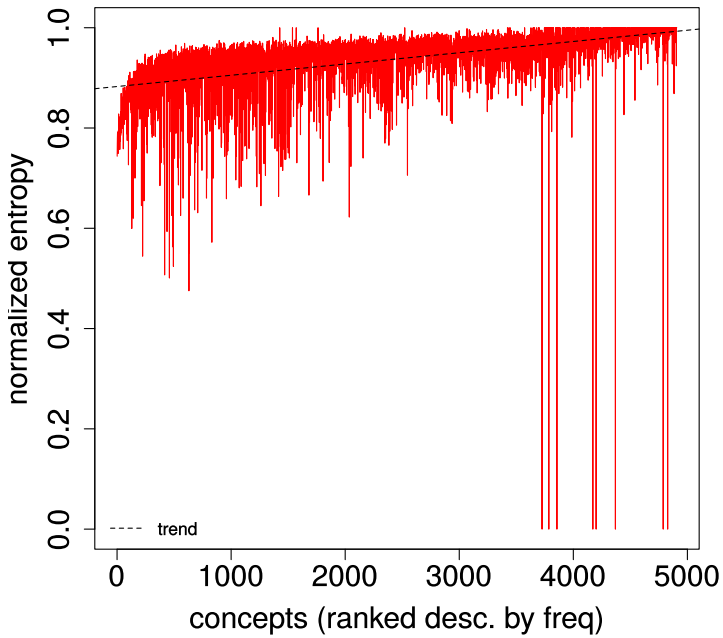
$$H(d) = -\frac{1}{n}\sum_{w_i}^{n} P(w_i) \cdot \log_2 P(w_i)$$

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

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

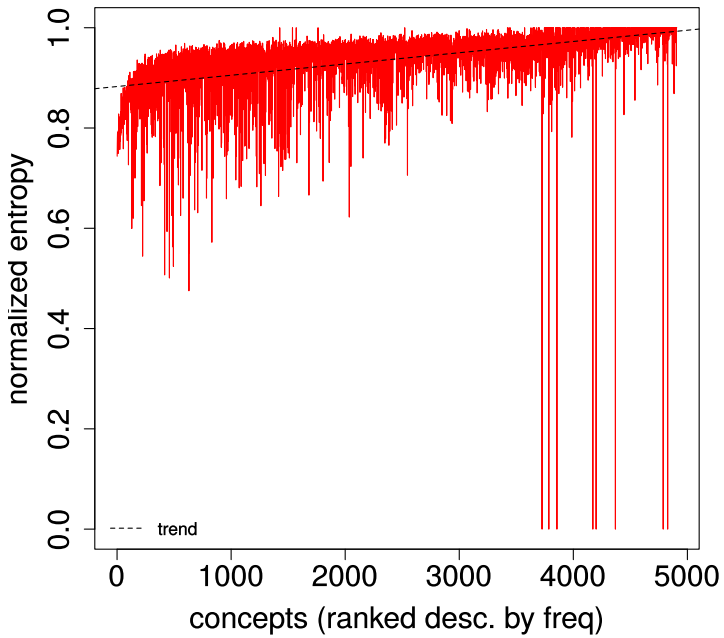
Quantitative: How specific are diseases?

Quantitative: How specific are diseases?

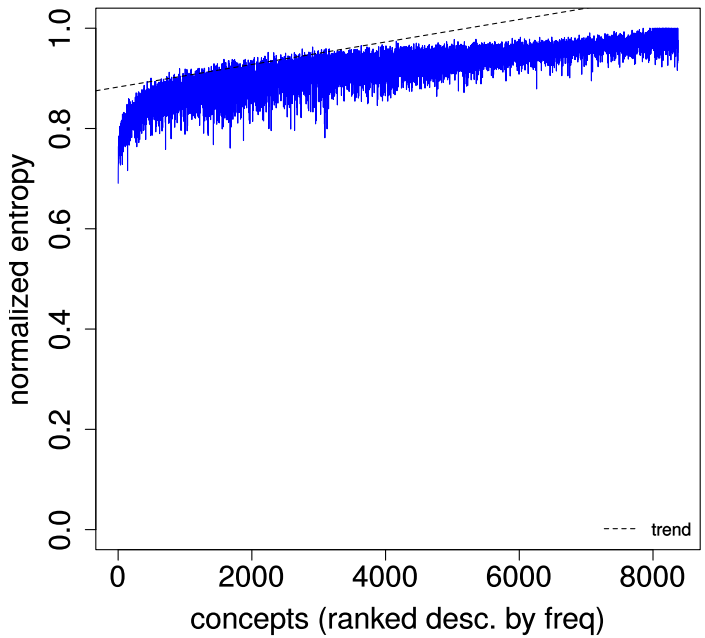


Twitter

Quantitative: How specific are diseases?

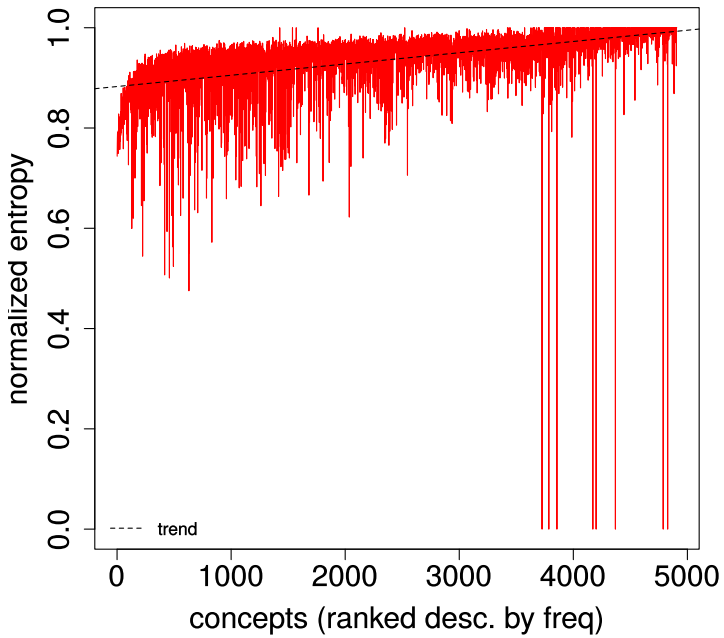


Twitter

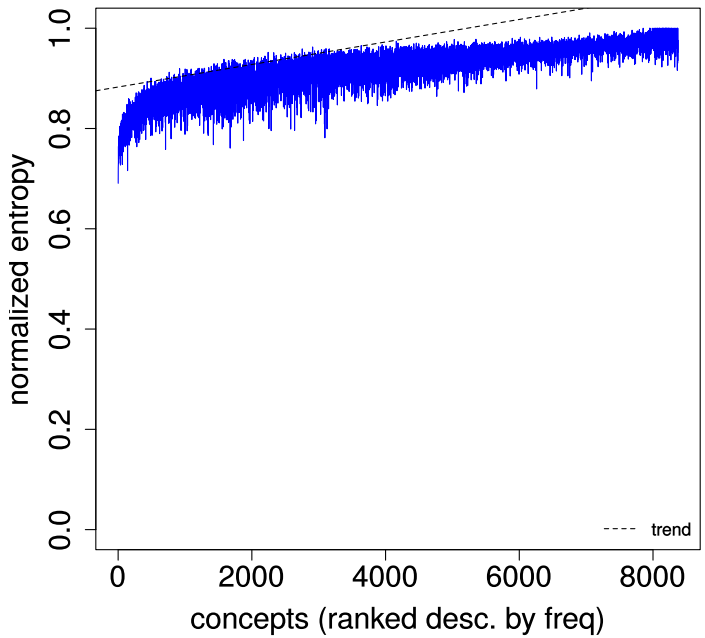


MEDLINE

Quantitative: How specific are diseases?



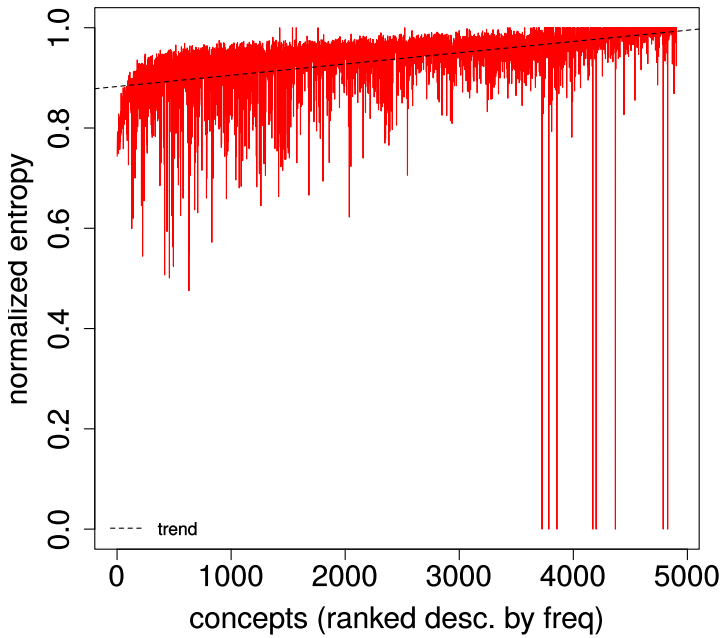
Twitter



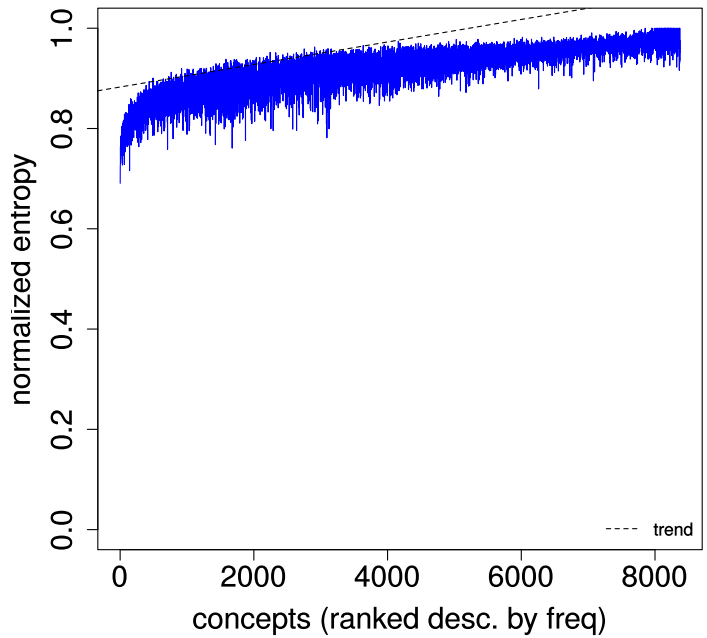
MEDLINE

- Frequent concepts have lower ambiguity (significant on both MEDLINE and Twitter)

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)?

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

- Use joint distributional model across both corpora

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

- Use joint distributional model across both corpora
- Similarity of two concepts/synonyms: Cosine similiarity

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

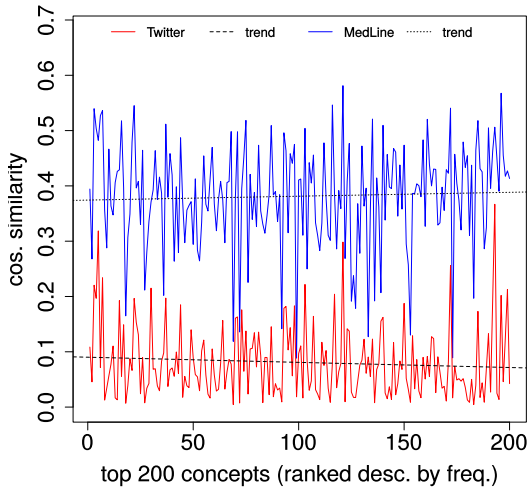
- Use joint distributional model across both corpora
- Similarity of two concepts/synonyms: Cosine similiarity
- Analyse Similarity between top 200 concepts

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

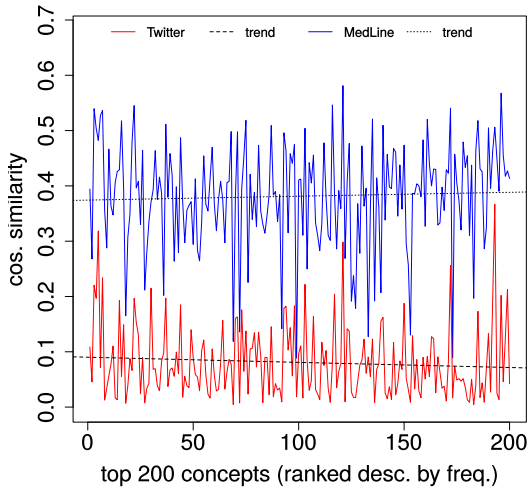
Similarity Analysis - Results

Similarity Analysis - Results

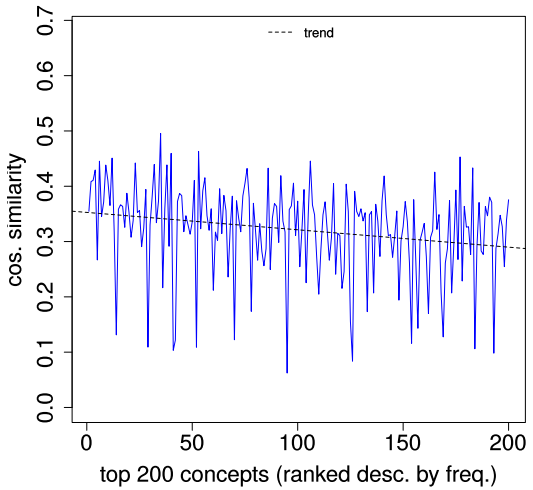


Intra-concept
Synonym Similarity

Similarity Analysis - Results

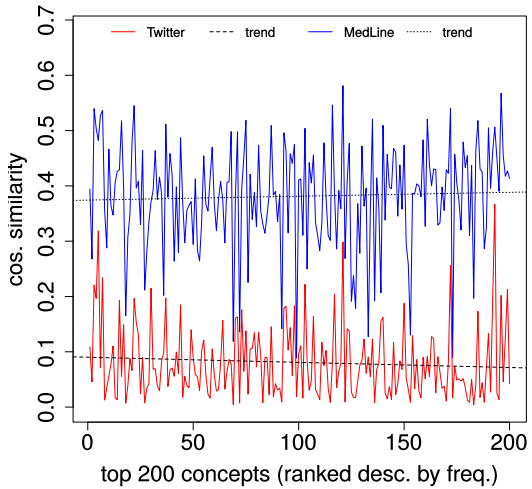


Intra-concept
Synonym Similarity

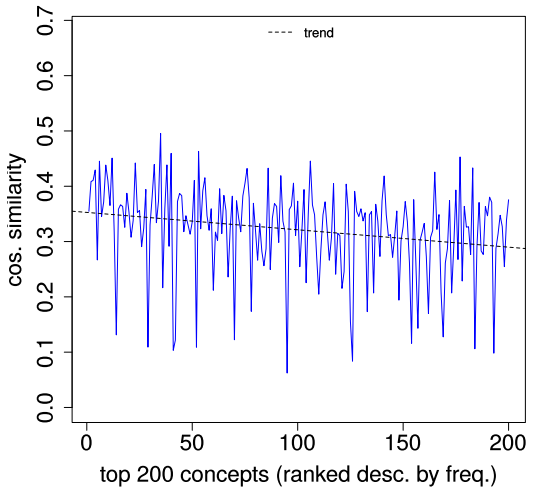


Concept Similarity

Similarity Analysis - Results



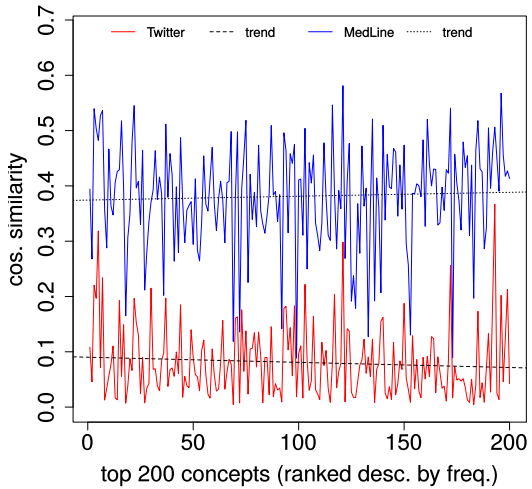
Intra-concept Synonym Similarity



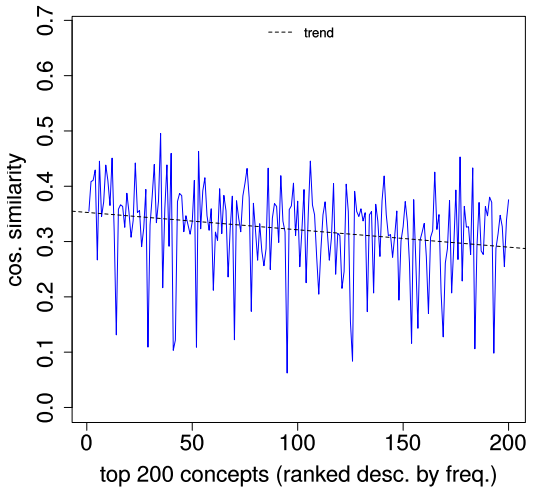
Concept Similarity

- Intra-concept similarity stable (avg.) w.r.t. frequency

Similarity Analysis - Results



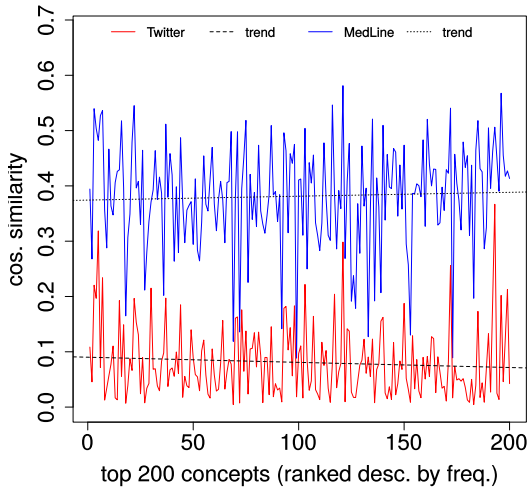
Intra-concept Synonym Similarity



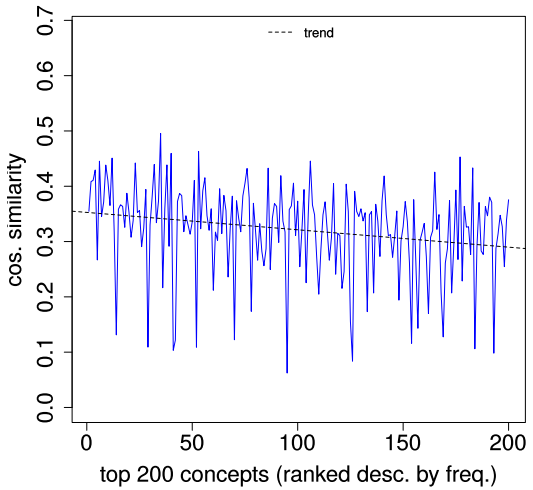
Concept Similarity

- Intra-concept similarity stable (avg.) w.r.t. frequency
- Intra-concept similarity higher on MEDLINE than in Twitter

Similarity Analysis - Results



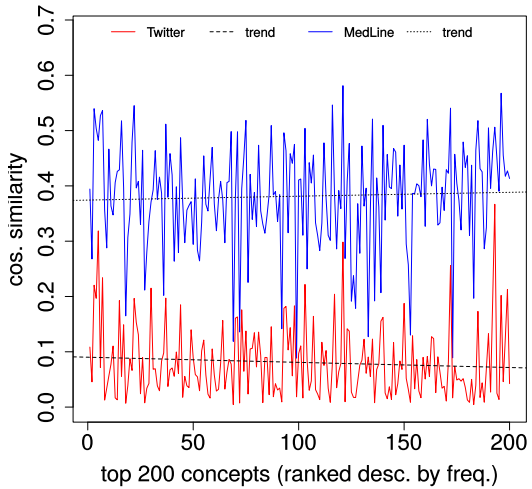
Intra-concept Synonym Similarity



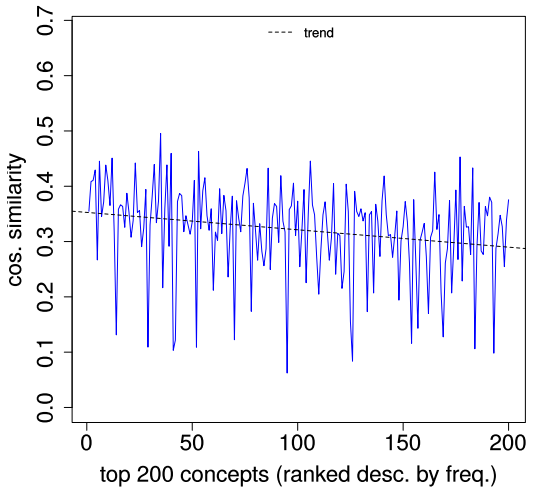
Concept Similarity

- 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.)

Similarity Analysis - Results



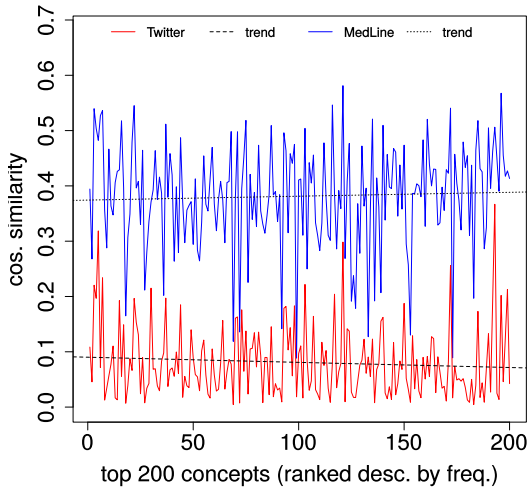
Intra-concept Synonym Similarity



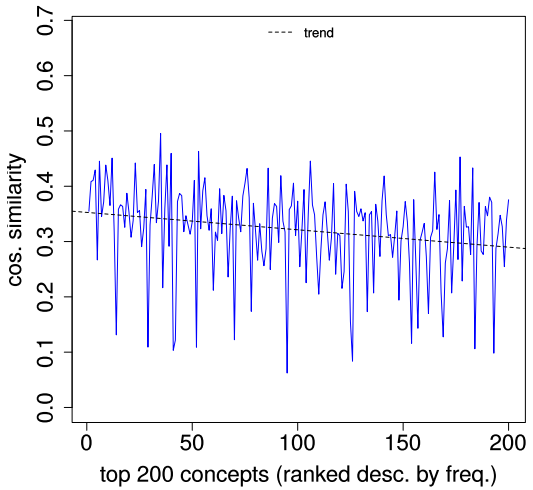
Concept Similarity

- 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

Similarity Analysis - 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 Analysis: Most/least similar concepts

Qualitative Analysis: 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

Qualitative Analysis: 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 Analysis: Synonym Level

(a) 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

(b) 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

Qualitative Analysis: Synonym Level

(a) 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

(b) 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

- Distributional analysis of disease concepts/synonyms in Twitter and MEDLINE
- Measured and compared distributional ambiguity – specificity – and similarity of disease concepts/synonyms
- Low distributional similarity among both corpora, coupled with higher ambiguity in Twitter compared to MEDLINE.
- Standard disease recognition methods such as DNorm result in high numbers of false positives
- Reason: larger use of catchphrases and metaphorical expressions in Twitter
- Future work: to separate non-disease name mentions from actual disease mentions in social media


Thank
You!

References I

- Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL 2014*, 2014.
- Georgiana Dinu, Nghia The Pham, and Marco Baroni. DISSECT - DIStributional SEmantics Composition Toolkit. In *Proceedings of ACL 2013*, 2013.
- Z. Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

References II

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu.
DNorm: Disease name normalization with pairwise learning to rank.
Bioinformatics, 29(22):2909–2917, 2013.

I. Dan Melamed.
Measuring semantic entropy.
In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics*, 1997.

References III


Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel E. Ginn, and Graciela Gonzalez.
Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features.
JAMIA, 22(3):671–681, 2015.