On the Semantic Similarity of Disease Mentions in MEDLINE® and Twitter

Camilo Thorne and Roman Klinger

Institut für Maschinelle Sprachverarbeitung (IMS), University of Stuttgart
{camilo.thorne,roman.klinger}@ims.uni-stuttgart.de

Abstract. Social media mining is becoming an important technique to track the spread of infectious diseases and to understand specific needs of people affected by a medical condition. A common approach is to select a variety of synonyms for a disease derived from scientific literature to then retrieve social media posts for subsequent analysis. With this paper, we question the underlying assumption that user-generated text always makes use of such names, or assigns them the same meaning as in scientific literature. We analyze the most frequently used concepts in MEDLINE® for semantic similarity to Twitter use and compare their normalized entropy and cosine similarities based on a simple distributional model. We find that diseases are referred to in semantically different ways in both corpora, a difference that increases in inverse proportion to the frequency of the synonym, and of the commonness of the disease or condition. These results imply that, when sampling social media for disease-related micro-blogs, query expressions must be carefully chosen, and even more so for rarily mentioned diseases or conditions.

Keywords: Social Media Mining, Twitter, MEDLINE®, Disease Names

1 Introduction

Named entity recognition (NER) is a well-established task in biomedical information extraction. It covers a wide variety of entity classes that can be tackled, for instance: gene and protein names [16], chemical names [7], drug names [6], or disease names [3]. Diseases are specifically interesting for a number of healthcare information extraction tasks related to, *e. g.*, pharmacovigilance [8,12,14,17], where social media corpora need to be processed. A key goal in pharmacovigilance is to detect if a disease or condition is spawned by a particular drug or medication, by tracking the way it is mentioned in social media over time. Another important task is to determine which are the most salient diseases and disease names. One key assumption is that diseases are referred to, used and crucially *meant* in social media – hence, by laymen – in a manner similar to scientific literature. Hence, it is on the one hand sufficient to apply entity recognition models and methods trained on scientific text, and on the other hand to reuse scientific terminology to query for biomedical-related microblogs such as tweets [8].

Furthermore, it has been observed that language in social media is more metaphorical, more prone to typos and newly coined words than more formal texts [15]. For instance in Twitter, people variously refer to schizophrenia as “schizo”, “derangement”,

"bipolar disorder", "delusional disorder", and use "schizophrenia" in a metaphorical manner (unrelated to health) as in "business schizophrenia".

In this paper, we hypothesize by contrast that authors in social media (here, Twitter) use disease names in a manner different from scientific literature. This hypothesis has important consequences, that, to the best of our knowledge, until now have not been fully studied by biomedical social media mining literature: It implies that only some biomedical terms are useful for retrieving biomedical microblogs. It also implies that such terms need not necessarily be canonical terms, but rather less technical – though salient – synonyms.

To test our hypothesis, we seek to answer the following research question: *do the most frequent diseases used in* MEDLINE® *correspond in meaning and variety of use to those discussed in Twitter?* To this end, we study and compare the meaning of normalized, canonical disease mentions (henceforth: *concepts*) and of their different surface-level realizations (henceforth: *synonyms*), in large samples of Twitter (approx. 150M words) and MEDLINE® abstracts (approx. 1B words). To model and compare the meaning of disease concepts and synonyms we resort to *distributional semantics* [5]. The main tenet of distributional semantics (the so-called "distributional hypothesis") is that the meaning of a linguistic expression can be characterized or approximated by the company it keeps in corpora, *i. e.*, by the words with which it co-occurs. Such words are called the distributional *context* of a word or phrase: disease concepts and synonyms in the present study. Contexts thereafter induce word distributions and vector space representations for concepts and synonyms.

2 Methods and Data Collection

Data Collection We build our experiment on top of MEDLINE®[1] (a large bibliographic database covering abstracts of biomedical papers since the early 1950's) and Twitter. A central element of our work is DNorm [8], a disease named entity recognizer and normalizer to MeSH and OMIM. The Medical Subject Headings (MeSH) terminology [9] and Online Mendelian Inheritance in Man (OMIM) [4] are controlled vocabularies of canonical, unique disease names organized into taxonomies of categories. We first apply DNorm to the last ten years of MEDLINE® (Jan. 2007 to Dec. 2017). We then focus our study to those concepts which appear to be of highest relevance in MEDLINE® and search for postings in Twitter using the 20 most frequent synonyms associated to the most frequent 100 MeSH concepts (with the official Twitter API, between Dec. 2017 and Mar. 2018). Detailed corpus statistics are shown in Table 1.

Named Entity Recognition and Frequency. We observe disease synonyms d_s and concepts d_c (normalized MeSH identifiers) and (2) identify their span within each unit of text (resp., a tweet or an abstract). For each each d_m, for $m \in \{s, c\}$, *frequency* is measured. This is done in order to rank diseases by frequency and (1) compute the Jaccard (set) similarity at frequency rank $r \leq k$ for the topmost k concepts in MEDLINE® and Twitter respectively: $\text{sim}_{\text{jacc}}(M_r, T_r) = |M_r \cap T_r|/|M_r \cup T_r|$. We also (2) test for concept frequency correlation across both corpora.

[1] https://www.nlm.nih.gov/bsd/pmresources.html

Table 1: Statistics of the samples studied. "Concepts" refer to MeSH *canonical* names.

Corpus	#Tokens	Units	Concepts	Synonyms	Time span
MEDLINE®	1,037,482,692	5,374,700 abstracts	8,386	2,190,522	Jan. 2007–Dec. 2017
Twitter	145,793,358	7,193,077 tweets	4,908	201,712	Dec. 2017–Mar. 2018

Specificity. We represent each d_m as a *distributional context* $\boldsymbol{D}_m$ – the bag or multiset of words with which d_m co-occurs in corpus C – by counting in a window of five words before and after the mention. Distributional contexts allow to quantify the semantic *specificity* (conversely, *ambiguity*) of d_m in each corpus via *normalized entropy*:

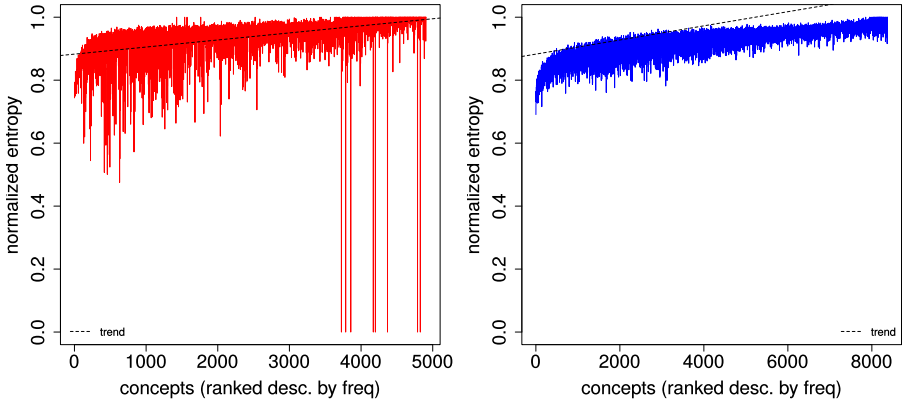
$$H_n(d_m) = -\sum_{\boldsymbol{D}_m(w)>0} \frac{P(w)\cdot \log_2 P(w)}{\log_2 D_m} \tag{1}$$

where $P(w)$ is estimated via relative frequencies, and $D_m = \sum\{\boldsymbol{D}_m(w) \mid \boldsymbol{D}_m(w) > 0\}$ is the *size* of $\boldsymbol{D}_m$. We use normalized entropy to be able to compute comparable measures (scaled to $[0, 1]$) for each d_m independent from context sizes. The higher $H_n(d_m)$, the more ambiguous d_m [10]. Thereafter we test (1) if concept frequency correlates with normalized entropy within each corpus, and (2) if normalized entropy in Twitter correlates with normalized entropy in MEDLINE®. Hence, contexts of co-occurring words can be seen as discrete word distributions. Entropy thus measures the dispersion of this induced distribution.

Similarity. We exploit distributional contexts to build a distributional model (vector space) $\mathbb{R}^{|W|}$, where each d_m is represented as a $|W|$-dimensional *distributional vector* $\boldsymbol{d}_m$ of log-scaled co-occurrence counts over the vocabulary W of MEDLINE® and Twitter. We rely on [2] to build the model. Classical count models are more appropriate in the context of this study than more state-of-the-art methods such as Word2Vec [11] or GloVE [13] word embeddings, which encode words directly into real-valued vectors and make it more difficult to compute entropies – that rely on discrete co-occurence counts [1]. Once we have learned the model from the distributional contexts, we compute cosine similarity for (1) the topmost k matching concepts in MEDLINE® and Twitter, *i. e.*, the d_cs in $M_k \cap T_k$. Additionally, we (2) study their *similarity spread*: we group their 20 most frequent synonyms to compute and compare their average similarities.

MeSH Hierarchy. MeSH IDs – disease concepts d_c – can be organized into a hierarchy $\boldsymbol{H}$ – a tree – of disease categories[2], from more general (root) to more specific (leaves). As an additional experiment, the relationship between concept normalized entropy in MEDLINE® and Twitter, and its position in $\boldsymbol{H}$ is studied. Highly ambiguous, high-entropy concepts should be located closer to the root node of the category hierarchy and have a low depth. To this end, we measure the depth of each concept d_c occurring in either Twitter or MEDLINE® in $\boldsymbol{H}$, and test if this measure correlates to their normalized entropy in either corpus.

[2] `https://www.nlm.nih.gov/mesh/intro_trees.html`



(a) Twitter (b) MEDLINE®

Fig. 1: Normalized entropies ranked by frequency in descending order. For each concept, we take as distributional context the union of those of all its synonyms.

3 Results

3.1 Experiments

Concept Specificity Analysis. In this experiment, we analyze if disease concepts mentioned in Tweets are more specific than concepts mentioned in MEDLINE®. Results are summarized in Figure 1. We observe a difference in concept normalized entropy distribution among both corpora. As Figure 1 shows, normalized entropy in Twitter is on average higher than in MEDLINE®: 0.95 vs. 0.92, the difference being statistically significant (t test with $p < 0.01$). Also, larger variations in normalized entropy can be seen in Twitter.

In both cases we observe a statistically significant negative correlation (-0.69 for MEDLINE® and -0.53 for Twitter, for the Kendall τ test for rank-sensitive correlation) between normalized entropy and frequency within each corpus. In Twitter moreover, disease concepts with both high normalized entropy and very low frequency can be observed, such as *Susceptibility of Obsesity* (OMIM UID 602025, point at frequency rank 4609 in Twitter) or *Pseudopseudohypoparathyroidism* (MeSH UID D011556, point at frequency rank 3713 in Twitter).

This seems due to the fact that frequent MEDLINE® concepts tend to be referred to with a smaller number of salient semantically specific synonyms, whereas tweets are overall more ambiguous. A cross-corpus correlation analysis (Kendall τ test) of both normalized entropy and frequency shows no significant correlation between normalized entropy or frequency distributions for concepts between Twitter and MEDLINE®.

Concept Similarity Analysis. In this experiment, we study the semantic similarity of MEDLINE® and Twitter, by computing the cosine similarity of the vector representations of diseases concepts and their most frequent synonyms.

Table 2: Jaccard and cosine similarity of t 100 concepts.

Frequency rank	Jaccard	Cosine (avg.)
top 20	0.212	0.365
top 40	0.356	0.358
top 60	0.446	0.345
top 80	0.553	0.342
top 100	0.550	0.338

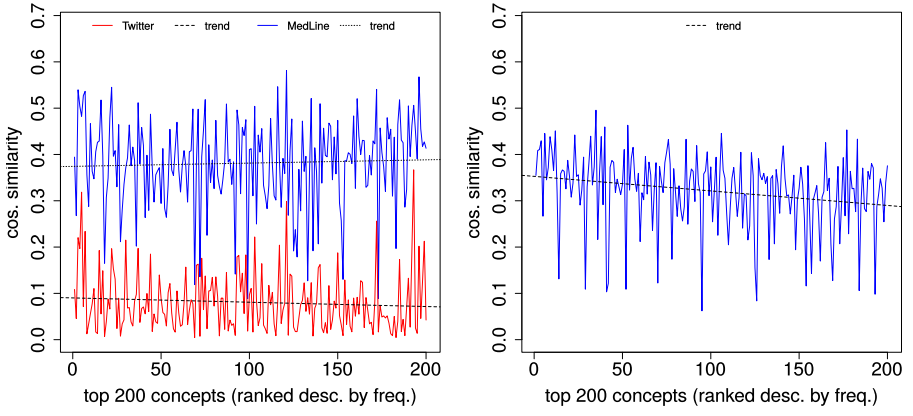
This is a key aspect of our analysis, in which we endeavor this time to determine and quantify the semantic (di)similarity of diseases in MEDLINE® and Twitter, with standard and robust distributional semantic techniques. The results are summarized by Tables 2, 3 and 4, and by Figure 2. The results show that, while there is actually a large overlap between the topmost 100 disease concepts mentioned both in Twitter and MEDLINE® (55% overlap/Jaccard similarity, see Table 2), distributional meanings differ considerably, yielding instead a 34% average cosine similarity for the same 100 concepts.

Interestingly (see Tables 3 and 4) common diseases such as *Hepatitis C* obtain a comparatively high cosine score, whereas very rare diseases such as *Behcet Syndrome* (a rare blood vessel chronic inflammation) obtain very low scores, as do diseases for which only ambiguous names exist, such as *BMD* (which may stand for *Becker muscular dystrophy*, or for bone mineral density, a mere symptom).

Figure 2b visualizes how similarity varies across matching concepts, ranked decreasingly by their frequency in MEDLINE®. As the reader can observe in the Figures, cross-corpus similarity has a slight tendency to decrease with frequency ($\tau = 0.18$, $p < 0.01$).

This seems to be due to differences in language use in both corpora: (1) different synonyms are associated to each concept, (2) they are assigned different meanings with (3) a higher semantic variability for Twitter compared to MEDLINE® – a fact consistent with our specificity analysis. Observations (2) and (3) are substantiated by Figure 2a that visualizes the average similarity among the (topmost 20) synonyms of each concept in each corpus, and shows that they are much higher in MEDLINE® than in Twitter. It is also possible to observe that concepts with above average intra-concept similarity in one corpus, exhibit an above average similarity in the other corpus as well. In fact, one can observe a statistically significant positive correlation ($\tau = 0.37$, $p < 0.01$).

Hierarchy Analysis. In this experiment, we study if specificity of disease concepts in Twitter and MEDLINE® decreases the higher the concepts are located in the MeSH disease category tree. We observe a slight, though statistically significant negative correlation, between normalized entropy and the depth of d_c in $\boldsymbol{H}$, *i. e.*, the lower the depth, the higher the specificity of d_c ($\tau = -0.129$ ($p < 0.01$) for MEDLINE®, -0.051 ($p < 0.01$) for Twitter). This can be interpreted as meaning that the distributional ambiguity as measured by $H_n(d_c)$ overlaps with the semantic generality of a disease concept d_c, while remaining largely distinct.



(a) Intra-concept Synonym Similarity

(b) Concept Similarity

Fig. 2: Similarity of concepts and synonyms, ranked by frequency in descending order. Tables 5 and 6 zoom into at concepts at rank 121 and 176 respectively.

3.2 Synonym Analysis

Our results are further substantiated by a qualitative analysis of concepts d_c and their synonyms d_s in both corpora, illustrated by Tables 5 and 6. We observe that higher similarity tend to coincide with a higher number of shared synonyms: in general the top two or three most frequent d_s for each d_c, more or less coincide across both corpora, but diverge afterwards the more dissimilar d_c in between Twitter and MEDLINE®. In Tables 5 and 6 we outline the main synonyms of an above average similarity disease concept, *Multiple Myeloma* (MeSH ID D009101, 0.39 similarity), and a below average similarity concept, *Angelman Syndrome* (MeSH ID D017204, 0.17 similarity). As the reader can see in the tables, while two out of three of the topmost synonyms of *Multiple Myeloma* coincide, the same only holds for one synonym for *Angelman Syndrome*.

Table 3: 7 most similar MeSH concepts.

MeSH ID	Similarity	Canonical name
D006526	0.496	Hepatitis C
D005910	0.463	Glioma
D003920	0.459	Diabetes Mellitus
D006521	0.453	Chronic Hepatitis
D000860	0.451	Hypoxia
D003327	0.446	Coronary Disease
D015658	0.445	HIV Infections
...	...	...

Table 4: 6 least similar MeSH concepts.

MeSH ID	Similarity	Canonical name
...	...	...
D015458	0.170	T Cell Leukemia
D002547	0.155	Cerebral Palsy
C536528	0.122	Van der Woude syndrome
C535984	0.116	Congenital bilateral aplasia of vas deferens
D029461	0.109	Sialic Acid Storage Disease
C537666	0.109	BMD

Table 5: 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

Table 6: 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

We also observe that a large number of false positives (false disease entities) are detected for Twitter. For instance, "Happiness", and "happiness" are detected in Twitter as synonyms of *Angelman Syndrome*, but also (in the 38th position) the catchphrase "AS IF IT". This likely explains the low synonym pairwise similarity observed in Figure 2, left, and the higher average normalized entropy observed in Twitter. The reason for this behavior is likely the strong bias of disease NER and normalization systems such as DNorm towards scientific terminology. Indeed, "happiness" may indicate a symptom of *Angelman Syndrome*, that is a severe genetic disorder affecting children and associated to frequent smiling. But in Table 6, a large share of those synonyms actually refer to emotions or wishes (as in "Happy birthday to our bright dancer (...) May he find happiness (...)").

4 Conclusions

We have carried out an extensive distributional analysis of the semantics of disease concepts and their synonyms in both social media (Twitter) and biomedical literature (MEDLINE®). To this end, we have measured and compared the normalized entropy – the distributional ambiguity – and distributional similarity of disease concepts observed in both corpora. Our analysis shows low distributional similarity among both corpora, coupled with a higher ambiguity in Twitter compared to MEDLINE®.

Our preliminary qualitative analysis shows that standard disease recognition methods such as DNorm result in high numbers of false positives, due to the larger use of catchphrases and metaphorical expressions in Twitter.

Future work will focus on the development of methods which can separate such non-disease name mentions from actual disease mentions in social media, and that help

in identifying tweets substantially similar to scientific texts. Ultimately, our goal is to build upon such methods to design techniques that identify and match disease-centric relations across both social media and MEDLINE®.

Acknowledgments. This work was supported by a grant from the Ministry of Science, Research and Arts of Baden-Württemberg to Roman Klinger.

References

1. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of ACL 2014 (2014)
2. Dinu, G., Pham, N.T., Baroni, M.: DISSECT - DIStributional SEmantics Composition Toolkit. In: Proceedings of ACL 2013 (2013)
3. Doğan, R.I., Leaman, R., Lu, Z.: NCBI disease corpus: a resource for disease name recognition and concept normalization. Journal of Biomedical Informatics 47, 1–10 (2014)
4. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A.: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research 33, 514–517 (2005)
5. Harris, Z.: Distributional structure. Word 10(23), 146–162 (1954)
6. He, L., Yang, Z., Lin, H., Li, Y.: Drug name recognition in biomedical texts: a machine-learning-based method. Drug Discovery Today 19(5), 610–617 (2014)
7. Klinger, R., Kolik, C., Fluck, J., Hofmann-Apitius, M., Friedrich, C.M.: Detection of IUPAC and IUPAC-like chemical names. Bioinformatics 24(13), 268–276 (2008)
8. Leaman, R., Islamaj Doan, R., Lu, Z.: DNorm: Disease name normalization with pairwise learning to rank. Bioinformatics 29(22), 2909–2917 (2013)
9. Lipscomb, C.E.: Medical Subject Headings (MeSH). Bulletin of the Medical Library Association 88(3), 265–266 (2000)
10. Melamed, I.D.: Measuring semantic entropy. In: Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics (1997)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
12. Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R.E., Gonzalez, G.: Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. JAMIA 22(3), 671–681 (2015)
13. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of EMNLP 2014 (2014)
14. Sarker, A., O'Connor, K., Ginn, R., Scotch, M., Smith, K., Malone, D., Gonzalez, G.: Social media mining for toxicovigilance: Automatic monitoring of prescription medication abuse from Twitter. Drug Safety 39(3), 231–240 (2016)
15. Seargeant, P., Tagg, C. (eds.): The Language of Social Media. Palgrave Macmillan, London (2014)
16. Wei, C.H., Kao, H.Y., Lu, Z.: GNormPlus: An integrative approach for tagging genes, gene families, and protein domains. BioMed Research International 2015(2015), ID 918710 (2015)
17. Yang, C.C., Yang, H., Jiang, L., Zhang, M.: Social media mining for drug safety signal detection. In: Proceedings of the 2012 International Workshop on Smart Health and Wellbeing (SHB 2012) (2012)