Mining Health-related Cause–Effect Statements with High Precision at Large Scale

Ferdinand Schlatt [⋆] **Dieter Bettin** [‡] **Matthias Hagen** [⋆] **Benno Stein** [†] **Martin Potthast** [§]

[⋆]Martin-Luther-Universität Halle-Wittenberg [†]Bauhaus-Universität Weimar
[‡]Westfälische Wilhelms-Universität Münster [§]Universität Leipzig

Abstract

An efficient assessment of the health relatedness of text passages is important to mine the web at scale to conduct health sociological analyses or to develop a health search engine. We propose a new efficient and effective termhood score for predicting the health relatedness of phrases and sentences, which achieves 69% recall at over 90% precision on a web dataset with cause–effect statements. It is more effective than state-of-the-art medical entity linkers and as effective but much faster than BERT-based approaches. Using our method, we compile the Webis Health CauseNet 2022, a new resource of 7.8 million health-related cause–effect statements such as "Studies show that stress induces insomnia" in which the cause ('stress') and effect ('insomnia') are labeled.

1 Introduction

Health sociology studies the interaction of society with health. An important subject is how consumers obtain and perceive health information. Since the web and search engines are among the most important sources, the quality of online health information has been studied so frequently in recent decades that three systematic reviews have been conducted (see Table 1). However, many of the individual studies address only a single medical condition and almost all were conducted manually. On average, only about 50–100 web pages were analyzed per study, sometimes only a single handpicked one, and never more than 1,524 pages.

Studying larger portions of the online health domain requires the automation of various acquisition tasks: (1) the discovery of health-related websites and web pages, (2) the identification and extraction of health-related statements from these pages, and (3) the attribution of health-related statements to authoritative sources (e.g., for fact-checking). While the first and third steps have been and continue to be the subject of ongoing research, the second step has received much less attention.

Systematic Review	Studies	Websites / Web Pages				
		Min	Max	Mean	Stddev	Total
Eysenbach et al. (2002)	79	3	1,147	100.5	157.7	7,836
Zhang et al. (2015)	165	3	388	78.5	73.4	12,870
Daraz et al. (2018)	157	1	1,524	50.3	133.9	7,891

Table 1: Sizes of systematic reviews of online health information studies. Some studies are part of multiple reviews; most do not distinguish websites and web pages.

To minimize manual data cleansing, we view the extraction of health-related statements as a precision-oriented task. And since a significant portion of consumers' health information needs ask about causes and effects (Bondarenko et al., 2022), we focus on health-related cause–effect statements (e.g., 'smoking causes cancer'). Extracting cause–effect statements in general has been thoroughly investigated in the past and several approaches extract them from web text efficiently and effectively (Yang et al., 2022). However, the extracted statements are usually not assigned to a specific domain.

Our contributions are: (1) A new approach for a high-precision assessment of a phrase's health relatedness (Section 3), which is more effective than state-of-the-art medical entity linkers and on par with BERT-based models but far more efficient (Section 4). (2) The Webis Health CauseNet 2022, a web-scale resource of health-related cause–effect statements (Section 5).[1]

2 Related Work

The impact of online health information on consumers has attracted the interest of the health sociology research community. For example, user surveys examine how consumers perceive online health information (Diaz et al., 2002), e-health services (Andreassen et al., 2007), or the quality of online health information (Sun et al., 2019).

[1]All our code and data to reproduce the experiments as well as the resource are publicly available under a permissive license: https://github.com/webis-de/COLING-22

Information quality appears to be the most studied characteristic from a health sociology perspective. Numerous studies systematically analyzed the quality of websites related to specific topics such as orthodontics (Jiang, 2000) or performance-enhancing drugs (Brennan et al., 2013), but more general studies with limitations to specific parts of the web are also common. Examples include studies of dietary advice (Cooper et al., 2012) or the misinterpretation (Yavchitz et al., 2012) and exaggeration (Sumner et al., 2014) of clinical trial results in online news. Recent studies also targeted web search snippets (Bondarenko et al., 2021) and social media (Suarez-Lledo and Alvarez-Galvez, 2021), particularly health misinformation on Twitter (Broniatowski et al., 2018; Bal et al., 2020).

To identify health-related web content, most previous work has focused on whole-page classification, e.g., using medical vocabularies to classify news articles (Watters et al., 2002; Zheng et al., 2002) or using convolutional neural networks for Reddit posts (Gkotsis et al., 2017). But there has been little work on classifying shorter passages of text as health-related (e.g., phrases), although this would allow for more fine-grained analyses at the statement level rather than at the whole-page level. Keyword extraction and automatic ontology creation with the goal of extracting prototypical words for a given domain are perhaps the most closely related tasks. For example, the C-value/NC-value method uses term frequencies to extract multi-word domain terms from a corpus (Frantzi et al., 2000). Its reliance on syntactic sentence structure, however, renders it inapplicable at the phrase level.

Contrastive termhood scores, which relate term frequencies from a domain corpus to frequencies from one or more out-of-domain corpora, can be applied more straightforwardly. These include $tf \cdot idf$-inspired measures (Basili et al., 2001; Kim et al., 2009), measures estimating how exclusive a term is for a domain (Ahmad et al., 1999; Park et al., 2008), and combinations or extensions thereof (Wong et al., 2007; Bonin et al., 2010). We transfer contrastive termhood scoring to measuring health relatedness of phrases (but also sentences) and compare it with the medical entity linkers cTakes (Savova et al., 2010), MetaMap (Aronson, 2001), QuickUMLS (Soldaini and Goharian, 2016), and ScispaCy (Neumann et al., 2019), as well as BERT-based classifiers (Devlin et al., 2019).

3 Measuring Health Relatedness

Assessing whether a phrase is health-related can be difficult without context; in particular for homonymous (same surface form, different meaning) or polysemous (same surface form, different sense) words. For instance, 'cancer' may refer to a health-related malignant tumor, but also to the zodiac sign, which is unlikely to appear in a health-related context. As such, the task of assessing a phrase's health relatedness can be viewed as an extension of word-sense disambiguation. Instead of the sense of a particular word, the domain of the sense of a phrase needs to be determined.

Since there are no large-scale labeled datasets for health relatedness assessment, we rely on contrastive termhood scores that use distant supervision to measure the degree of a word or concept being specific to a certain domain (Kageura and Umino, 1996). Instead of training on explicitly labeled words or phrases, contrastive termhood scores are trained on texts that are heuristically labeled. Specifically, a word's or phrase's domain specificity depends on its "prominence" in domain-specific or out-of-domain corpora. In this section, we discuss three existing contrastive termhood scores and then explain how we adapt and apply them to health relatedness assessment.

3.1 Existing Contrastive Termhood Scores

The termhood scores contrastive weight (CW) (Basili et al., 2001), term domain specificity (TDS) (Park et al., 2008), and discriminative weight (DW) (Wong et al., 2007) rely on a corpus H of domain-specific texts (in our case: health-related texts) and at least one contrastive corpus G of general or out-of-domain texts (in our case: Wikipedia). To score a term t (a word or phrase), CW, TDS, and DW use occurrence frequencies: the absolute corpus occurrence frequency $freq_C(t)$ (i.e., the absolute number of occurrences of t in corpus C), the relative corpus occurrence frequency $rel_C(t) = freq_C(t)/|C|$ (where $|C|$ denotes some appropriate variant of corpus size like number of words or n-gram occurrences), and the inverse corpora frequency $icf(t)$ defined for H and G together as

$$icf(t) = \log\left(\frac{|H| + |G|}{freq_H(t) + freq_G(t)}\right).$$

The contrastive weight CW of a term t is similar to $tf \cdot idf$ but uses the corpus-oriented frequencies:

$$\mathrm{CW}(t) = \log\left(freq_H(t) + 1\right) \cdot icf(t).$$

For the term domain specificity TDS, we unify the slightly different definitions of Ahmad et al., Park et al., and Wong et al. as

$$\mathrm{TDS}(t) = \log\left(\frac{rel_H(t)+1}{rel_G(t)+1}+1\right).$$

Finally, the discriminative weight DW was originally defined as the product of CW and a version of TDS that uses $freq_C(t)$ instead of $rel_C(t)$. Since the values of such an "unnormalized" TDS depend on corpus size, we use our above corpus-agnostic normalized version but still compute DW as

$$\mathrm{DW}(t) = \mathrm{CW}(t) \cdot \mathrm{TDS}(t).$$

3.2 Our Generalized Termhood Scores

The above termhood scores were originally meant to help augment taxonomy vocabularies or to find terms missing in a dictionary for some specific domain. As such, the input terms are assumed to be rather short and quite domain-related noun phrases from which the "best" scoring ones are to-be-added to the vocabulary. In pilot experiments for our case of assessing the health relatedness of phrases from the web, we observed that phrases like 'fracture at the base of the skull' receive very low termhood scores even though they are clearly health-related. The reason is that such longer phrases as "a whole" have quite low occurrence frequencies even in medical corpora. A first straightforward idea could be to average a phrase's individual word's termhood scores. However, for the above example with many out-of-domain or stop words, this also does not work well. We thus propose two schemes that improve on the simple average and on the original termhood scores' treatment of longer phrases as "a whole". In our schemes, we also enable the assessment to prioritize precision or recall.

Our first scheme uses a weighted average of a phrase's individual word's termhood scores (i.e., their individual health relatedness) to assess the health relatedness of a phrase. The idea is that by giving words that have a high individual health relatedness score a higher weight, more health-related phrases can be found (i.e., improved recall). Similarly, by giving words that have a low individual health relatedness score a higher weight, the precision of the assessment on the phrase level should improve. Formally, in our first scheme, we compute the termhood score of an m-word phrase as the generalized mean of the word's individual termhood scores $x_1, \ldots, x_m$:

$$M_r(x_1, \ldots, x_m) = \left(\frac{1}{m}\sum_{i=1}^{m} x_i^r\right)^{\frac{1}{r}},$$

where r is a real-valued parameter. For $r = 1$, the generalized mean corresponds to the arithmetic mean. By increasing r, the mean is biased towards the higher-valued termhood scores and vice versa. In the extreme cases of $M_{-\infty}$ or M_{∞}, the minimum or maximum x_i is returned. At the limit for r approaching 0, the generalized mean corresponds to the geometric mean (Jensen, 1998).

As our second scheme, we propose to also compute the weighted average termhood over the n-grams of a phrase. For instance, while the unigrams 'risk' and 'factor' are relatively unrelated to health, the bigram 'risk factor' certainly is health-related. The above generalized mean scheme already increases the termhood of 'risk factor' compared to a simple average but the high occurrence frequency of the bigram itself is an even better indicator of its health relatedness. Since longer n-grams usually have quite low occurrence frequencies, especially in smaller corpora, we only average the termhood scores of a phrase's uni-, bi-, or trigrams with a parameter n determining the length of the longest used n-grams. For example, the generalized score of the phrase 'cancer risk factor' based on the termhood score $s(.)$ (could be CW, TDS, or DW) with $n = 2$ then is the average of $M_r(s(\text{'cancer risk'}), s(\text{'risk factor'}))$ and $M_r(s(\text{'cancer'}), s(\text{'risk'}), s(\text{'factor'}))$.

Corpus	Language	Documents	Words
Wikipedia	mixed, layperson	12,265,374	$3.0{\cdot}10^9$
PubMed	scientific, abstracts	31,847,923	$3.8{\cdot}10^9$
PubMed Centr.	scientific, full texts	3,611,361	$5.4{\cdot}10^9$
Textbook	clinical, educational	434	$1.4{\cdot}10^7$
Encyclopedia	clinical, layperson	67,967	$9.3{\cdot}10^6$

Table 2: Characteristics of the contrastive and health-related corpora used for the termhood scores.

3.3 Contrastive and Health Domain Corpora

Table 2 shows basic characteristics of our employed corpora. We select Wikipedia[2] as our contrastive general corpus G since it covers a wide variety of domains and is easily accessible. As candidates for health corpora H, we consider and evaluate four alternatives, each with its own (dis)advantages.

[2]Dump of all English Wikipedia articles from July 1, 2021.

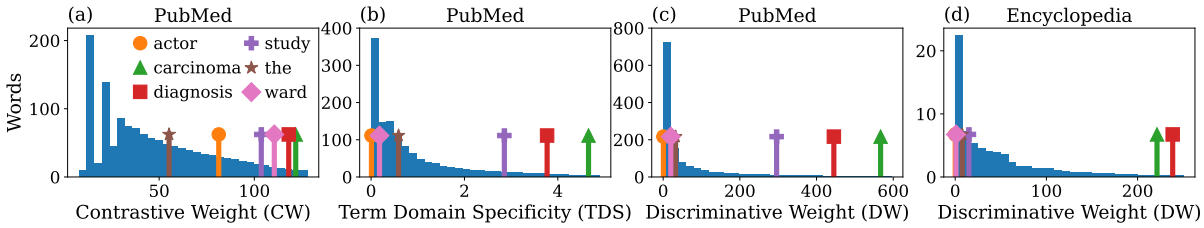

Figure 1: Termhood score frequencies of (a) CW, (b) TDS, and (c) DW on the PubMed corpus, and of (d) DW on the Encyclopedia corpus. Number of words in thousands (from the respective corpus). Example words are highlighted.

The first three health corpora use documents provided by the National Library of Medicine: (1) a dump of over 30 million MEDLINE abstracts from PubMed,[3] (2) a subset of over 3 million full-text publications from PubMed Central,[4] and (3) 434 textbooks from the textbook and monograph category of the NCBI Bookshelf.[5] While both PubMed-based corpora are large scale, their language is mainly scientific. The textbook corpus contains more clinical language, which we hypothesize to more closely match the language of health-related phrases on arbitrary web pages.

Finally, as our fourth health corpus, we crawled the entries of five consumer-oriented medical online encyclopedias.[6–10] Since they are written in layperson's terms, we assume their language to be most similar to the target language distribution used for health-related phrases on web pages.

3.4 Pilot Inspection and Comparison

To get a first impression of the scores and the corpus impact, we inspect the unigram termhood score distributions in general and for some example words. Figures 1 (a–c) show the score distributions of CW, TSD, and DW with PubMed as the domain-specific corpus H. Apparently, all scores rank the shown example out-of-domain words and the stop word lower than the shown example health-related words. However, the assessment of CW and TSD can differ substantially for specific terms. For example, 'ward' occurs frequently within texts from the PubMed corpus so that CW attributes a rather high health relatedness. At the same time, 'ward' also occurs frequently in the general domain and the lacking "exclusiveness" leads TDS to score it relatively low. Unsurprisingly, the product score DW amplifies the extremes of both scores.

As for the effect of different health corpora, Figures 1 (c) and (d) contrast the DW scores using the PubMed corpus (rather scientific language) to using the Encyclopedia corpus (rather layperson language). As an example result, the word 'study' has a comparably high termhood score using the PubMed corpus, but is ranked like a non-health-related word using the Encyclopedia corpus.

[3] https://pubmed.ncbi.nlm.nih.gov/
[4] https://www.ncbi.nlm.nih.gov/pmc/
[5] https://www.ncbi.nlm.nih.gov/books
[6] http://health.am/encyclopedia
[7] https://medlineplus.gov/encyclopedia.html
[8] https://merriam-webster.com/medical
[9] https://ucsfhealth.org (various subpages)
[10] https://www.rxlist.com/drug-medical-dictionary/article.htm

Dataset	Text Type	Health	Length	Size
CauseNet-F-Phrase	Phrase pairs	21.4%	7.2	1,000
CauseNet-P-Phrase	Phrase pairs	50.3%	3.4	1,000
CauseNet-F-Sentence	Sentences	22.4%	30.3	1,000

Table 3: Characteristics of our three annotated datasets, including the ratio of health-related entries, the average number of words per cause–effect phrase pair or sentence, and the number of entries. CauseNet-F: sampled from CauseNet-Full, CauseNet-P: sampled from CauseNet-Precision.

4 Evaluation

In this section, we evaluate our generalized termhood method on datasets of cause–effect statements and compare it to state-of-the-art medical entity linkers and BERT-based approaches.

4.1 Annotated Datasets

Table 3 depicts the general characteristics of three cause–effect datasets we sampled from the web-scale CauseNet resource (Heindorf et al., 2020), a graph of over 11 million cause–effect pairs (e.g., 'stress → insomnia') extracted from the ClueWeb12.[11] The CauseNet extraction used a two-stage approach: (1) candidate sentences were gathered using a set of lexico-syntactic patterns representing causal language, and (2) a BiLSTM-CRF model extracted cause–effect pairs from the

[11] https://www.lemurproject.org/clueweb12/

candidates. Obviously, a pair may be extracted from multiple different sentences (e.g., 'stress → insomnia' from 'stress causes insomnia' or from 'insomnia can be a result of stress'). All the sentences from which a specific pair is extracted are the *support* of that pair.

CauseNet comes in different versions: CauseNet-Full and CauseNet-Precision. In CauseNet-Full, all the extracted cause–effect pairs are contained, while CauseNet-Precision only contains pairs that were extracted by at least two different lexico-syntactic extraction patterns. The idea is that pairs supported by sentences from more patterns are less likely to be false positive causal statements.

From each CauseNet version, we randomly sampled 1,000 cause–effect pairs for our CauseNet-F-Phrase and CauseNet-P-Phrase datasets (F: full, P: precision). Note that the pairs sampled from CauseNet-Precision typically are shorter (3.4 vs. 7.2 words; cf. Table 3)—shorter phrases are more likely to be extracted by more than one pattern.

To label the health relatedness of the cause–effect pairs, we had three annotators who first labeled a kappa-test subset of 100 samples from CauseNet-F-Phrase. The achieved agreement was substantial (Cohen's kappa of 0.76), so that after a discussion of the disagreement cases, the annotators then each independently labeled disjoint thirds of the remaining CauseNet-F/P-Phrase data. Interestingly, according to the labeling, the pairs sampled from CauseNet-Precision are much more health related (50.3% vs. 21.4%; cf. Table 3).

Finally, for each of the 1,000 pairs in CauseNet-F-Phrase, we randomly selected one of the supporting sentences to complement the phrase-based dataset with a sentence-based dataset. The respective CauseNet-F-Sentence set then was also labeled with respect to health-relatedness by our three annotators. Indeed, for eleven causal phrase pairs that were labeled as non-health-related and for one pair that was labeled as health-related, the sampled supporting sentence then was labeled in the opposite way since it added important context. For example, the pair 'sound → particular feeling' was labeled as non-health related but the corresponding sentence 'Any sound that is related to trauma can trigger a particular feeling.' was labeled as health-related due to the explicit connection to trauma.

We randomly split each of the three datasets to use 80% as the training data for parameter tuning and the remaining 20% as the actual test set.

4.2 Medical Entity Linkers

We compare our generalized termhood method to the state-of-the-art medical entity linkers cTakes, MetaMap, QuickUMLS, and ScispaCy. To determine the health relatedness of a (multi-word) term by applying one of these entity linkers, we use the proportion of words that the linker matches to medical concepts in some background knowledge base like the UMLS Metathesaurus (Humphreys and Lindberg, 1993).

More formally, as the entity linking-based health relatedness of a term t, we use the the ratio $|e|/|\hat{t}|$, where $|e|$ denotes the length (in words) of the substring of t that the linker detects as mentions of some entity, and $|\hat{t}|$ denotes the length of t without the stopwords[12] that are not contained in any entity mention detected by the respective linker.

As the background knowledge base for the entity linkers we try four different options: (1) the full UMLS (a mix of medical vocabularies of varying specificity), (2) the combined UMLS subsets RxNorm and SNOMED CT (more specific clinical vocabulary), (3) UMLS restricted to the 21 most frequent semantic types (ST21pv subset) as proposed in the MedMentions entity linking dataset (Mohan and Li, 2019), and (4) the combined RxNorm and SNOMED CT restricted to ST21pv.

4.3 BERT-Based Approaches

Besides medical entity linkers, we also compare our generalized termhood method to BERT-based classifiers fine-tuned to predict whether a sequence of tokens is health-related. To test the effect of domain specific embeddings, we compare classifiers based on pre-trained BERT (Devlin et al., 2019), SciBERT (Beltagy et al., 2019), and PubMedBERT (Gu et al., 2022) models. To further fine-tune these models to the task, we first construct additional training datasets where noun phrases (or sentences) from Wikipedia are paired with noun phrases (or sentences) from the PubMed or the Encyclopedia corpus. To align with the evaluation datasets, we extract noun phrases and sentences using spaCy.[13] All models were fine-tuned with a batch size of 32 and a learning rate of $5 \cdot 10^{-5}$ for 100,000 steps on one NVIDIA A100 GPU. Due to the large corpora sizes, we used early stopping and halted training when no decrease in training loss was recorded for 15 consecutive samples (taken every 1,000 steps).

[12] English nltk stop words list.
[13] https://spacy.io/

4.4 Assessing the Health Relatedness of Cause–Effect Pairs

To assess the health relatedness of a cause–effect pair, we combine an approach's individual scores for the cause and effect phrase into one score based on which a decision about the health relatedness can be made (i.e., whether it is above some threshold). Just like in our generalized termhood scores, we use the generalized mean as the combination scheme. The extreme case of $M_{-\infty}$ then corresponds to an `AND` combination (cause and effect score need to exceed a decision threshold), while M_{∞} corresponds to an `OR` combination (cause or effect above some threshold suffice). These setups can thus be interpreted as precision or recall-oriented, respectively.

4.5 Assessing the Health Relatedness of Cause–Effect Sentences

To assess the health relatedness of a complete sentence, we basically use the different approaches as if the input was a phrase. The entity linking-based methods link the entities in the sentence and then compute the same ratio $|e|/|\hat{t}|$ but with $\hat{t}$ now being the sentence without non-linked stopwords. The BERT-based approaches also get the sentence as input and classify it as a whole as health-related or not. And also our generalized termhood scores simply treat an input sentence like a phrase.

4.6 Parameters and Optimization Criteria

We explore different hyperparameter values of the approaches in grid searches on the 80% training sets of our annotated cause–effect datasets (the other 20% are the test sets). For the termhood method, we experiment with four health corpora (PubMed, PubMed Central, Textbook, Encyclopedia) and a maximum n-gram length $n \in \{1, 2, 3\}$. For the generalized mean M_r, we try $r \in \{0, \pm 1, \pm 2, \pm 5, \pm 10, \pm\infty\}$ for combining the n-gram scores, as well as for combining the cause and effect scores of all approaches.

For the entity linkers, we explore linking against the full UMLS with either all semantic types or just the ST21pv subset, or linking just against the combined RxNorm and SNOMED CT subset with either all types or just the ST21pv subset. For the QuickUMLS and ScispaCy linkers, we try similarity thresholds in steps of 0.1 between $[0.7, 1.0]$ and $[0.6, 0.9]$, respectively, and explore small (*sm*) and large (*lg*) models for ScispaCy. Finally, the BERT-based models are fine-tuned in four variations: on sentences or on phrases from the PubMed or from the Encyclopedia corpus.

We conduct three grid searches for three different optimization scenarios. In the first two scenarios, we optimize for precision or recall and identify a parameterization and decision threshold with the best recall (or precision) that can be achieved at an operating point of a precision (or recall) of 0.9. For example, in the scenario of precision optimization, we only consider parameter combinations from the grid search and decision thresholds that achieve a precision of at least 0.9 on the training set. From these, we then only consider the parameterizations with the highest recall and from these, we select one with the highest precision. In the third optimization scenario, we target the Matthews correlation coefficient (MCC). The MCC combines the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as

$$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP{+}FP)(TP{+}FN)(TN{+}FP)(TN{+}FN)}}.$$

Since MCC generates a high score only when the majority of the positive instances and the majority of the negative instances are classified correctly, it is regarded as one of the best ways to derive one score from a binary classifier's confusion matrix (Chicco and Jurman, 2020).

4.7 Results

Table 4 shows the effectiveness of the different approaches' best parameterizations from the training sets run on the test sets of CauseNet-F-Phrase, CauseNet-P-Phrase, and CauseNet-F-Sentence. The BERT-based approaches are usually the most effective. Our generalized termhood methods often are slightly less effective but the difference to the best BERT-based approach is hardly ever statistically significant (bootstrapping with 100,000 permutations, $p < 0.05$, Bonferroni-corrected). The entity linking-based approaches, though, almost always are significantly less effective than the best BERT-based approach.

While being almost as effective as the best BERT-based approach, our generalized termhood method is substantially more efficient (see Table 5). On an AMD EPYC 7F72 processor, even without parallelization, the generalized termhood is up to 107 times faster on phrases and up to 47 times faster on the longer sentences than the BERT models. By

	Approach	Precision Optimized				Recall Optimized				MCC Optimized					
		Parameters	M_r	P	R	Parameters	M_r	P	R	Parameters	M_r	P	R	F1	M
CauseNet-F-Phrase	cTakes	RS ST21pv	M_{-2}	0.00	0.00†	RS	M_2	0.28†	0.95	RS	M_5	0.43	0.67	0.52	0.39†
	MetaMap	RS ST21pv	$M_{-\infty}$	0.67	0.05†	RS	M_2	0.35†	0.95	RS	M_2	0.52	0.82	0.63	0.54†
	QuickUMLS	RS ST21pv, 0.8	M_{-1}	0.00	0.00†	RS, 0.8	M_2	0.30†	0.87	RS, 0.8	M_2	0.57	0.41	0.48	0.38†
	ScispaCy	RS, *sm*, 0.9	$M_{-\infty}$	0.00	0.00†	UMLS, *lg*, 0.6	M_2	0.26†	0.79	RS, *sm*, 0.6	M_{10}	0.35	0.64	0.45	0.29†
	BERT	ENC, NP	M_1	0.91	0.77	ENC, NP	M_5	**0.76**	**0.97**	ENC, NP	M_2	0.81	0.90	0.85	**0.82**
	SciBERT	ENC, NP	M_5	0.94	0.77	ENC, NP	M_2	**0.76**	0.95	ENC, NP	M_2	0.82	0.85	0.84	0.80
	PMBERT	ENC, NP	M_2	0.94	**0.82**	ENC, NP	M_1	0.73	0.95	ENC, NP	M_2	0.80	**0.92**	**0.86**	**0.82**
	CW	ENC, n=2, M_2	M_{-1}	0.85	0.72	ENC, n=3, M_5	M_2	0.58†	0.92	ENC, n=1, M_5	M_2	**0.84**	0.79	0.82	0.77
	TDS	ENC, n=2, M_1	M_1	**1.00**	0.69	ENC, n=2, M_0	M_{-1}	0.67	0.92	ENC, n=3, M_0	M_1	0.76	0.82	0.79	0.74
	DW	ENC, n=1, M_1	M_1	**1.00**	0.74	ENC, n=3, M_1	M_∞	0.62	0.90	ENC, n=2, M_2	M_1	0.78	0.90	0.83	0.79
CauseNet-P-Phrase	cTakes	RS	M_1	0.79	0.34†	RS	M_2	0.63†	**0.96**	RS	M_{-2}	0.75	0.56	0.64	0.33†
	MetaMap	RS	M_{-5}	0.76	0.85	RS	M_{-2}	0.69†	0.94	RS	M_{-5}	0.76	0.85	0.80	0.52†
	QuickUMLS	RS, 1.0	M_{10}	0.83	0.46†	RS, 0.9	M_2	0.66†	0.88	RS, 0.8	M_1	0.75	0.79	0.77	0.46†
	ScispaCy	RS, *sm*, 0.7	$M_{-\infty}$	0.71	0.33†	UMLS, *lg*, 0.7	M_2	0.60†	0.88	RS, *sm*, 0.8	M_{10}	0.66	0.80	0.73	0.30†
	BERT	ENC, NP	M_5	0.91	0.88	ENC, NP	M_{10}	0.92	0.91	ENC, NP	M_5	0.91	0.88	0.90	0.77
	SciBERT	ENC, NP	M_{10}	**0.97**	**0.89**	ENC, NP	M_5	0.90	0.90	ENC, NP	M_5	0.90	0.90	0.90	0.78
	PMBERT	ENC, NP	M_5	0.96	0.88	ENC, NP	M_2	0.91	0.95	ENC, NP	M_5	**0.96**	0.88	**0.92**	**0.83**
	CW	ENC, n=1, M_{10}	M_5	0.95	0.72	ENC, n=1, M_{10}	M_2	0.80†	0.91	ENC, n=1, M_5	M_2	0.83	0.89	0.86	0.66
	TDS	ENC, n=3, M_1	M_1	0.93	**0.89**	ENC, n=2, M_1	M_1	**0.93**	0.89	ENC, n=2, M_1	M_1	0.89	**0.93**	0.91	0.79
	DW	ENC, n=2, M_1	M_1	0.92	0.88	ENC, n=3, M_{10}	M_1	**0.93**	0.91	ENC, n=3, M_5	M_1	0.90	0.91	0.91	0.79
CauseNet-F-Sentence	cTakes	RS	–	0.70	0.20†	RS	–	0.25†	**0.91**	RS	–	0.57	0.46	0.51	0.42†
	MetaMap	RS	–	**1.00**	0.03†	RS	–	0.29†	**0.91**	RS	–	0.42	0.49	0.45	0.33†
	QuickUMLS	RS ST21pv, 1.0	–	**1.00**	0.03†	RS, 1.0	–	0.32†	**0.91**	RS, 1.0	–	0.49	0.49	0.49	0.38†
	ScispaCy	RS, *lg*, 0.8	–	0.88	0.20†	RS, *sm*, 0.9	–	0.25†	0.89	RS, *lg*, 0.6	–	0.42	0.60	0.49	0.37†
	BERT	ENC, NP	–	0.86	0.51	ENC, NP	–	0.59	0.83	ENC, NP	–	0.76	0.74	**0.75**	**0.70**
	SciBERT	ENC, NP	–	0.89	0.49	ENC, NP	–	0.53	0.89	ENC, NP	–	0.64	0.66	0.65	0.57
	PMBERT	ENC, NP	–	0.83	**0.57**	ENC, NP	–	0.59	0.86	ENC, NP	–	**0.87**	0.57	0.69	0.66
	CW	ENC, n=3, M_5	–	0.75	0.43	ENC, n=1, M_{10}	–	0.56	0.89	ENC, n=1, M_5	–	0.67	0.63	0.65	0.58
	TDS	ENC, n=1, M_1	–	0.84	0.46	ENC, n=1, M_2	–	**0.62**	**0.91**	ENC, n=2, M_0	–	0.69	0.63	0.66	0.59
	DW	ENC, n=3, M_1	–	0.83	0.43	ENC, n=2, M_2	–	0.59	0.83	ENC, n=2, M_1	–	0.71	**0.77**	0.74	0.68

Table 4: Effectiveness on the test sets as precision (P), recall (R), F1, or Matthews correlation coefficient (MCC) of the best parameterization of each approach optimized for precision, recall, or MCC on the respective training set. The operating point for precision / recall optimization is set to 0.9 on the training data (gray scores indicate that 0.9 could not be reached during training). Statistically significant differences to the best approach for a dataset and optimization criterion (best scores highlighted in bold) are denoted by † ($p < 0.05$, Bonferroni-corrected for the nine comparisons in each group). For entity linkers, the usage of UMLS or combined RxNorm and SNOMED CT vocabulary (RS) restricted / or not to the ST21pv subsets, and, where applicable, similarity thresholds or spaCy model size (*sm* or *lg*) are reported. BERT models were fine-tuned on the PubMed (PM) or Encyclopedia (ENC) corpus using sentences (S) or noun phrases (NP). Termhood scores use either the PubMed (PM), PubMed Central (PMC), Textbook (TB), or Encyclopedia (ENC) corpus, a maximum n-gram size of n, and the generalized mean M_r. In the phrase scenarios, the generalized mean M_r for combining the cause and effect scores is also reported.

precomputing the n-gram frequencies and then parallelizing hash table lookups in the inference phase, the termhood scores could be even further sped up. As for memory efficiency, the n-gram frequencies and the BERT model checkpoints have a similar memory footprint of about 400MB.

As for the assessment of phrases vs. sentences, our results in Table 4 show that most approaches are substantially more effective on just the phrase pairs than on sentences (e.g., BERT-based: 0.12 to 0.23 better MCC on CauseNet-F-Phrase than on CauseNet-F-Sentence; termhood: 0.11 to 0.19 better MCC). Only the ScispaCy entity linking approach is really more effective on sentences than phrases (MCC improves by 0.08).

Entity Linking-based Approaches The entity linking-based approaches mostly use the combined RxNorm and SNOMED CT vocabulary (some-

Approach	Phrase		Sentence	
	ms	Speedup	ms	Speedup
cTakes	119.68	0.5	212.12	0.2
MetaMap	49.64	1.2	120.28	0.4
QuickUMLS	7.23	8.3	8.98	5.3
ScispaCy	16.38	3.7	15.96	3.0
PubMedBERT	60.19	1.0	47.77	1.0
Termhood n=1	**0.56**	**107.5**	**1.02**	**46.8**
Termhood n=2	0.93	64.7	1.97	24.2
Termhood n=3	1.27	47.4	2.82	16.9

Table 5: Run time efficiency of the different approaches' most effective parameterization on the CauseNet-F-Phrase and -Sentence test sets. Time per instance averaged over 10 runs, speedup computed against PubMedBERT as the most effective approach from Table 4.

times only the ST21pv subset) and only once the full UMLS. Still, even with such a "restricted" vocabulary, the entity linking-based approaches hardly achieve really high precision values—or only at the expense of very low recall. One reason is that even specifically tailored health-related entity vocabularies still contain many terms that are only loosely health-related and then yield false positive results on the cause–effect statements. Interestingly, only the MetaMap and the ScispaCy parameterizations "attempt" to compensate for this by using generalized averages M_r with $r < 0$ for the cause–effect combination when optimizing for precision. Still, the many 0-values for precision optimization on CauseNet-F-Phrase indicate that no health-related statements are found. This is again caused by a "precision problem". In the grid search on CauseNet-F-Phrase, all entity linking-based approaches achieve precision values of at least 0.9 but at a very tiny recall (6–8 true positives). On the test sets, these low-recall parameterizations do not detect any of the health-related statements.

Another drawback of some entity linkers in the phrase scenarios is their reliance on syntactic parsing to detect candidate mentions—the parses might not be too meaningful for (short) phrases. ScispaCy in particular relies on parsing and thus is less effective than the other linkers on phrases—even though ScispaCy is one of the best medical entity linkers in full text scenarios (Vashishth et al., 2021).

Overall, the entity linking-based approaches achieve rather low effectiveness compared to the other approaches and are also slower than our new generalized termhood method. In our scenario of assessing the health relatedness of phrases and sentences, they are not really a good option.

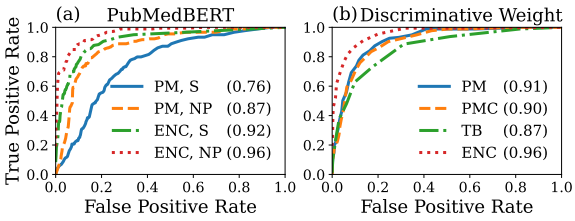

Figure 2: ROC curves and AUC values (in parentheses) (a) for fine-tuning PubMedBERT on the CauseNet-F-Sentence training data using sentences (S) or noun phrases (NP) from the PubMed (PM) or Encyclopedia (ENC) corpus, and (b) for discriminative weight DW ($n = 1, r = 1$) on the CauseNet-F-Phrase training data with the health corpora PubMed (PM), PubMed Central (PMC), Textbooks (TB), or Encyclopedia (ENC).

BERT-based Approaches Interestingly, domain specific pre-training only has a minor positive effect for the BERT-based models. While PubMedBERT often is the most effective, SciBERT and even the domain-agnostic BERT usually are almost as effective. Interestingly, all prefer the recall-oriented higher values of r in the generalized average M_r for the cause–effect score combination.

Another observation is that the best BERT-based approaches all are fine-tuned on noun phrases from the Encyclopedia corpus—even in the sentence scenario. As an example, Figure 2 (a) shows the ROC curves and AUC values for PubMedBERT in different fine-tuning setups on the CauseNet-F-Sentence training data. Fine-tuning on the Encyclopedia corpus clearly achieves better AUC values (at least 0.06 over PubMed) as does fine-tuning on phrases instead of sentences (at least 0.04).

Termhood-based Approaches Among the termhood-based approaches, TDS and DW are more effective than CW in almost all scenarios.

Further analysis of the termhood methods' parameters shows that the health corpus is very important: the best parameterizations all use the Encyclopedia corpus. Figure 2 (b) details this observation for DW. The AUC-ROC value with the Encyclopedia corpus is by far the best; the health corpus' fit to the target language is crucial. The n-gram length and the generalized mean setup are less important. Table 6 shows several ablation results. In most cases, the ablated setups achieve equal or lower effectiveness, but in some cases higher effectiveness. By fixing $n = 1$, the MCC decreases by 0.00 to 0.03 on most datasets (except TDS / CW on sentences where it increases). Fixing the phrase-internal averaging to the arithmetic mean (M_1) only decreases the MCC by a maximum of 0.05.

	Appr.	P Opt.		R Opt.		M Opt.	
		$n=1$	M_1	$n=1$	M_1	$n=1$	M_1
F-Phrase	CW	-0.10	-0.08	-0.03	-0.03	0.00	-0.05
	TDS	0.00	0.00	-0.05	0.03	-0.02	0.02
	DW	0.00	0.00	-0.04	0.00	-0.01	-0.01
P-Phrase	CW	0.00	0.00	0.00	0.00	0.00	-0.02
	TDS	-0.05	0.00	0.01	0.00	-0.03	0.00
	DW	-0.07	0.00	-0.03	0.00	-0.01	0.03
F-Sent.	CW	0.03	-0.03	0.00	0.00	0.00	0.05
	TDS	0.00	0.00	0.00	-0.11	0.07	0.10
	DW	-0.03	0.00	0.00	-0.03	0.02	0.00

Table 6: Ablation study indicating the difference in effectiveness on the test sets from the best parameterizations optimized for precision (P), recall (R), or the Matthews correlation coefficient (M) to the best parameterizations when ablating the n-gram length (fixed $n = 1$) or the generalized mean (fixed M_1 for combining the n-gram scores). Operating point for precision / recall is set to 0.9. The difference is given with respect to the "interesting" measure (i.e., drop in recall for the P columns, drop in precision for the R columns, and drop in MCC for the M columns).

Overall, our generalized termhood-based methods are much faster but not significantly less effective than the BERT-based approaches—with TDS and DW usually being better than CW. When applying the generalized termhood methods, it is crucial to choose a good health corpus while optimizing the other parameters (n-gram length, averaging) only leads to smaller improvements.

5 Webis Health CauseNet 2022

By applying our generalized termhood method for health relatedness assessment to the complete CauseNet, we create the new Webis Health CauseNet 2022 resource consisting of health-related cause–effect statements found on the web. It is important to note, that the statements in CauseNet—and thus also in our Webis Health CauseNet 2022—are only *claimed* cause–effect statements. For many of the contained statements, scientific evidence can surely be found (e.g., 'stress → insomnia') while for many other this might not be possible (e.g., 'incorrect placement of jupiter → diabetes'). Still, it could be interesting to analyze websites that contain many claimed health-related cause–effect statements with respect to whether medical evidence exists or not. Such health-sociological analyses are now enabled by our Webis Health CauseNet 2022 resource at web scale since the URLs of the pages from which a statement was extracted are part of CauseNet.

Subset	Statements	Sentences	P	R
Prec (P)	103,273	1,259,339	0.93	0.89
Prec (MCC)	112,707	1,340,873	0.89	0.93
Full (P)	2,201,071	5,680,635	1.00	0.74
Full (MCC)	3,206,964	7,842,464	0.78	0.90

Table 7: Characteristics of the four Webis Health CauseNet 2022 versions. Number of statements / supporting sentences, and estimated precision and recall of the precision- or MCC-optimized termhood extraction.

The Webis Health CauseNet 2022 comes in four different versions[14] based on the best termhood parameterizations from the evaluation. Table 7 contains some general characteristics. The two smaller versions are extracted from CauseNet-Prec (statements with high support) by using TDS optimized for precision or MCC, while the two larger versions are extracted from CauseNet-Full by using DW optimized for precision or MCC.

6 Conclusions

We have proposed generalized termhood-based methods that effectively and efficiently assess the health relatedness of phrases. On cause–effect statements from the web, our new approaches are almost as effective as the best BERT-based approaches while being much faster. Approaches using state-of-the-art medical entity linkers are slower and less effective. When configuring our new termhood-based methods, it is crucial to select a background health corpus that matches the target language distribution while the other parameters (n-gram length, averaging) are less important.

Using our methods, we have extracted the Webis Health CauseNet 2022 resource of health-related cause–effect statements from the web-scale CauseNet. Based on Webis Health CauseNet 2022, health-sociological analyses of online cause–effect relations are now possible at an unprecedented scale compared to previous small-scale analyses of health-related online information.

Finally, our termhood-based assessment could also be useful in retrieval scenarios. For instance, given the termhood scores' efficiency, they could directly be used at search engine side to quickly assess the health relatedness of some query not seen before (for other queries, the clicked documents usually suffice to assess the health relatedness) and to possibly adopt the retrieval accordingly (e.g., preferring medical resources).

[14] Available under a permissive license: https://github.com/webis-de/COLING-22

References

Khurshid Ahmad, Lee Gillam, and Lena Tostevin. 1999. University of Surrey participation in TREC8: Weirdness indexing for logical document extrapolation and retrieval (WILDER). In *Proceedings of The Eighth Text REtrieval Conference (TREC 1999), Gaithersburg, Maryland, USA, November 17–19, 1999*.

Hege K. Andreassen, Maria M. Bujnowska-Fedak, Catherine E. Chronaki, Roxana C. Dumitru, Iveta Pudule, Silvina Santana, Henning Voss, and Rolf Wynn. 2007. European citizens' use of e-health services: A study of seven countries. *BMC Public Health*, 7(1):53.

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proceedings of the American Medical Informatics Association Annual Symposium (AMIA 2001), Washington, DC, USA, November 3–7, 2001*.

Rakesh Bal, Sayan Sinha, Swastika Dutta, Rishabh Joshi, Sayan Ghosh, and Ritam Dutt. 2020. Analysing the extent of misinformation in cancer related tweets. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020), Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8–11, 2020*, pages 924–928.

Roberto Basili, Alessandro Moschitti, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. 2001. A contrastive approach to term extraction. In *Proceedings of Terminologie et Intelligence Artificielle (TIA 2001), Nancy, France, May 3–4, 2001*, pages 119–128.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), Hong Kong, China, November 3–7, 2019*, pages 3613–3618.

Alexander Bondarenko, Ekaterina Shirshakova, Marina Driker, Matthias Hagen, and Pavel Braslavski. 2021. Misbeliefs and biases in health-related searches. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM 2021), Virtual Event, Queensland, Australia, November 1–5, 2021*, pages 2894–2899.

Alexander Bondarenko, Magdalena Wolska, Stefan Heindorf, Lukas Blübaum, Axel-Cyrille Ngonga Ngomo, Benno Stein, Pavel Braslavski, Matthias Hagen, and Martin Potthast. 2022. A benchmark for causal question answering. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022), Gyeongju, Republic of Korea, October 12–17, 2022, (to appear)*.

Francesca Bonin, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2010. A contrastive approach to multi-word extraction from domain-specific corpora. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, May 17–23, 2010*.

Brian Brennan, Gen Kanayama, and Harrison Pope. 2013. Performance-enhancing drugs on the web: A growing public-health issue. *The American Journal on Addictions*, 22(2):158–161.

David A. Broniatowski, Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra Quinn, and Mark Dredze. 2018. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American Journal of Public Health*, 108(10):1378–1384.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21:6.

Benjamin E. J. Cooper, William E. Lee, Ben M. Goldacre, and Thomas A. B. Sanders. 2012. The quality of the evidence for dietary advice given in UK national newspapers. *Public Understanding of Science*, 21(6):664–673.

Lubna Daraz, Allison S. Morrow, Oscar J. Ponce, Wigdan Farah, Abdulrahman Katabi, Abdul Majzoub, Mohamed O. Seisa, Raed Benkhadra, Mouaz Alsawas, Prokop Larry, and M. Hassan Murad. 2018. Readability of online health information: A meta-narrative systematic review. *American Journal of Medical Quality*, 33(5):487–492.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019), Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Joseph A. Diaz, Rebecca A. Griffith, James J. Ng, Steven E. Reinert, Peter D. Friedmann, and Anne W. Moulton. 2002. Patients' use of the internet for medical information. *Journal of General Internal Medicine*, 17(3):180–185.

Gunther Eysenbach, John Powell, Oliver Kuss, and Eun-Ryoung Sa. 2002. Empirical studies assessing the quality of health information for consumers on the world wide web: A systematic review. *JAMA*, 287(20):2691–2700.

Katerina T. Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J. P. Hubbard, Richard J. B. Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using informed deep learning. *Scientific Reports*, 7(1):45141.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):2:1–2:23.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. CauseNet: Towards a causality graph extracted from the web. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM 2020), Virtual Event, Ireland, October 19–23, 2020*, pages 3023–3030.

Betsy L. Humphreys and Donald A. B. Lindberg. 1993. The UMLS project: Making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association*, 81(2):170–177.

Jerry L. Jensen. 1998. Some statistical properties of power averages for lognormal samples. *Water Resources Research*, 34(9):2415–2418.

You-Ling Jiang. 2000. Quality evaluation of orthodontic information on the world wide web. *American Journal of Orthodontics and Dentofacial Orthopedics*, 118(1):4–9.

Kyo Kageura and Bin Umino. 1996. Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2):259–289.

Su Nam Kim, Timothy Baldwin, and Min-Yen Kan. 2009. An unsupervised approach to domain-specific term extraction. In *Proceedings of the Australasian Language Technology Association Workshop (ALTA 2009), Sydney, Australia, December 3–4, 2009*, pages 94–98.

Sunil Mohan and Donghui Li. 2019. MedMentions: A large biomedical corpus annotated with UMLS concepts. In *Proceedings of the 1st Conference on Automated Knowledge Base Construction (AKBC 2019), Amherst, MA, USA, May 20–22, 2019*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task (BioNLP@ACL 2019), Florence, Italy, August 1, 2019*, pages 319–327.

Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C. Gates. 2008. An empirical analysis of word error rate and keyword error rate. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008), Brisbane, Australia, September 22–26, 2008*, pages 2070–2073.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Luca Soldaini and Nazli Goharian. 2016. QuickUMLS: A fast, unsupervised approach for medical concept extraction. In *Proceedings of the 2nd SIGIR Workshop on Medical Information Retrieval (MedIR 2016), Pisa, Italy, July 21, 2016*.

Victor Suarez-Lledo and Javier Alvarez-Galvez. 2021. Prevalence of health misinformation on social media: Systematic review. *Journal of Medical Internet Research*, 23(1):e17187.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A. Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy, and Christopher D. Chambers. 2014. The association between exaggeration in health related science news and academic press releases: Retrospective observational study. *BMJ*, 349:g7015.

Yalin Sun, Yan Zhang, Jacek Gwizdka, and Ciaran B. Trace. 2019. Consumer evaluation of the quality of online health information: Systematic literature review of relevant criteria and indicators. *Journal of Medical Internet Research*, 21(5):e12522.

Shikhar Vashishth, Denis Newman-Griffis, Rishabh Joshi, Ritam Dutt, and Carolyn P. Rosé. 2021. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *Journal of Biomedical Informatics*, 121:103880.

Carolyn R. Watters, Wanhong Zheng, and Evangelos E. Milios. 2002. Filtering for medical news items. In *Proceedings of the 65th ASIS&T Annual Meeting (ASIST 2002), Philadelphia, PA, USA, November 18–21, 2002*, volume 39, pages 284–291.

Wilson Wong, Wei Liu, and Mohammed Bennamoun. 2007. Determining termhood for learning domain ontologies using domain prevalence and tendency. In *Proceedings of the Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Queensland, Australia, December 3–4, 2007*, pages 47–54.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64(5):1161–1186.

Amélie Yavchitz, Isabelle Boutron, Aida Bafeta, Ibrahim Marroun, Pierre Charles, Jean Mantz, and Philippe Ravaud. 2012. Misrepresentation of randomized controlled trials in press releases and news coverage: A cohort study. *PLOS Medicine*, 9(9):e1001308.

Yan Zhang, Yalin Sun, and Bo Xie. 2015. Quality of health information for consumers on the web: A systematic review of indicators, criteria, tools, and evaluation results. *Journal of the Association for Information Science and Technology*, 66(10):2071–2084.

Wanhong Zheng, Evangelos E. Milios, and Carolyn R. Watters. 2002. Filtering for medical news items using a machine learning approach. In *Proceedings of the American Medical Informatics Association Annual Symposium (AMIA 2002), San Antonio, TX, USA, November 9–13, 2002*.