Matching Consumer Health Vocabulary with Professional Medical Terms Through Concept Embedding

Yue Wang[1], Jian Tang[2], V.G.Vinod Vydiswaran[1], Kai Zheng[3], Hua Xu[4], Qiaozhu Mei[1]

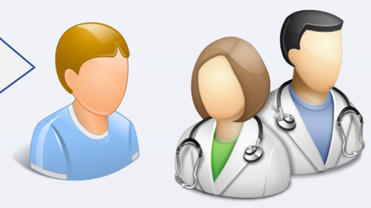
[1]University of Michigan; [2]HEC Montréal; [3]University of California, Irvine;

[4]University of Texas Health Science Center at Houston

MOTIVATION

❖ There exists language mismatch between consumers/laypersons and health care professionals (Zeng & Tse '06). For example:

- *loss of appetite*
- *pain killer*
- *heart attack*
- *kidney stones*
- *crank; ice; meth*
- …



- *anorexia*
- *analgesic*
- *myocardial infarction*
- *kidney calculi*
- *methamphetamine*
- …

❖ Bridging this language gap is an important step for:

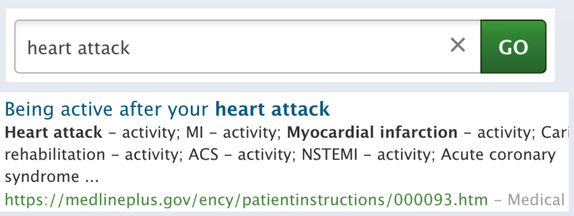
Consumers to search for relevant health information online

Consumers to make informed decisions

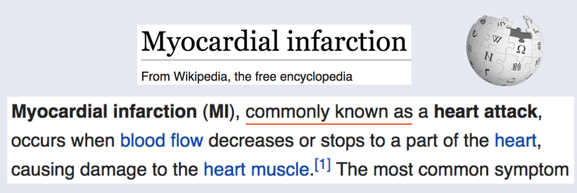
Professionals to communicate health information & share knowledge to consumers

❖ Two previous studies on mining consumer health vocabulary (CHV):

➢ Consumer search queries in MedlinePlus (Zeng & Tse '06):



➢ Community-generated Wikipedia entries, "also known as" or "commonly known as" (Vydiswaran et al '14):



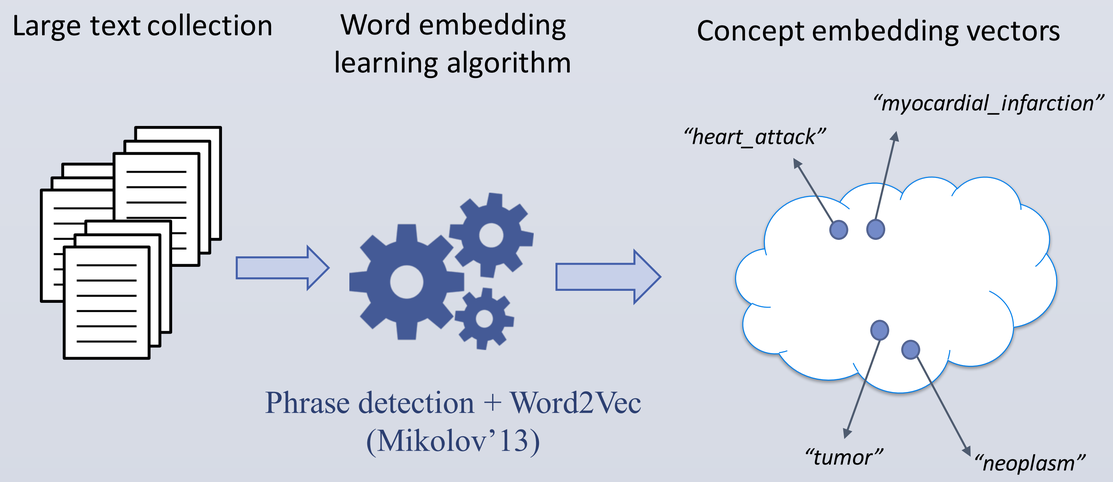
❖ These approaches obtain highly selective and accurate consumer-professional concept pairs, but many relevant pairs maybe missing, i.e. low coverage.

❖ **In this study, we aim at high coverage**: given *any* concept in consumer health vocabulary, how can we match it to professional medical terms (and vice versa)?

METHODS

❖ We learn **vector representations** for professional and consumer medical concepts from large-scale text corpus (100 dimensions).

❖ Similarity between two concepts is measured by the **cosine similarity** of their vectors.



❖ **Input**: three large-scale text collections:

- ➢ 2.2 million **MedHelp** posts (3 GB consumer-generated texts)
- ➢ 7 million **MEDLINE** abstracts (13 GB professional-generated texts)
- ➢ 4.7 million **Wikipedia** articles (11 GB open domain texts)

❖ **Output**: vector representation of medical concepts, including:

- ➢ 300K **words and phrases** from text collection
- ➢ 494K **medical concepts** from UMLS, including two consumer health vocabularies:
  - o CHV-1 (Vydiswaran et al '14), 892 consumer/professional concept pairs
  - o CHV-2 (Zeng & Tse '06), a random subset of 1,000 pairs
- ➢ Multi-word concept vector = average of word vectors

EVALUATION & RESULTS

❖ Given a consumer concept, we return a **ranked list of concepts by decreasing cosine similarity**, and see if the corresponding professional concept is ranked high. We use Mean Reciprocal Rank (MRR) to evaluate such ranked lists.

$$\text{Reciprocal Rank (RR)} = \frac{1}{hit\ rank} \qquad \text{(higher is better)}$$

Example: Query: heart attack

1. cardiac arrest
2. angina
3. myocardial infarction √

⇨ $RR = \frac{1}{3}$

Table 1. Mean reciprocal rank of corresponding consumer/professional concept.

	CHV-1 (Vydiswaran et al '14)		CHV-2 (Zeng & Tse '06)	
	Con ⇛ Pro	Pro ⇛ Con	Con ⇛ Pro	Pro ⇛ Con
Wikipedia	0.161	0.158	0.403	0.371
MedHelp	0.177	0.176	0.462	0.443
MEDLINE	0.192	0.190	0.485	0.459
MedHelp+MEDLINE	0.192	0.188	0.483	0.457

Con ⇛ Pro: use consumer concept as the query, compute the reciprocal rank of its corresponding professional concept. Pro ⇛ Con: vice versa.

Example ranked lists

Query: nocturnal_myoclonus
sleep myoclonus
sleep disorder hypersomnia
periodic limb movement sleep disorder
periodic leg movements sleep
night terrors sleep
periodic limb movement disorder
narcolepsy and hypersomnia
involuntary movement extrapyramidal restlessness
daytime somnolence sleep disorder

Query: seasonal affective disorder
seasonal affective disorder disease finding
bipolar affective disorder
unipolar depressive illness
manic depressive illness
bipolar affective disorder early onset
bipolar disorder manic depression
bipolar affective disorder atypical nos
generalized anxiety
major depressive disorder
(expected answer: winter blues)

Query: winter blues
summer depression ☜
blues laundry
winter wren
blues music
(expected answer: seasonable affective disorder)

Learned word/phrase neighborhoods

Query: analgesic	Query: painkiller	Query: myocardial_infarction	Query: heart_attack
analgetic	pain_reliever	infarction_mi	heartattack
opioid_analgesic	narcotic	acute_myocardial	heart-attack
pain-relieving	painkillers	myocardial_infraction	heart_attck
sedative	toradol	q-wave_myocardial	heart_attact
anti-emetic	paracetamol	infarction_ami	stroke
analgesics	codeine	reinfarction	cardiac_arrest
non-opioid_analgesic	analgesic	ami	angina
antihyperalgesic	pain_reliver	unstable_angina	heart_atttack
opioid-sparing	pain_killers	mi	heart_atack
narcotic_analgesic	nsaid	myocardial_infarctions	heart_attack-
analgesia	muscle_relaxer	q-wave_mi	massive_mi
antipyretic	morphine	infarction	sudden_death
analgesic_efficacy	tordal	non-q-wave_myocardial	stoke
narcotic	codiene	re-infarction	sudden_cardiac
antitussive	dilaudid	q-wave_infarction	clogged_arteries
antinociceptive	muscle_relaxant	non-q_wave	acute_mi
anti-nociceptive	pain-killer	non-q-wave_infarction	aheart_attack
non-opioid_analgesics	darvocet	postmyocardial_infarction	hart_attack
sedative_analgesic	panadeine	postinfarction_angina	pulmonary_embolism
antiemetic	codine	postinfarction	myocardial_infarction

❖ **Conclusion**: On average, a consumer or professional concept is 14th closest to its counterpart in CHV-1, and 7th closest to its counterpart in CHV-2.

❖ **Future work**: (1) label the concepts as consumer- or professional-oriented by propensity scores[2]; (2) invite medical experts for further review.

Acknowledgments This study is supported by NLM 2R01LM010681-05 and NSF IIS-1054199.

1. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. Journal of the American Medical Informatics Association. 2006;13(1):24–29.
2. Vydiswaran VV, Mei Q, Hanauer DA, Zheng K. Mining consumer health vocabulary from community-generated text. In: AMIA. vol. 2014. American Medical Informatics Association; 2014.
3. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Neural Information Processing Systems, 2013.

Contact: Yue Wang, raywang@umich.edu