Towards the Semantic Interpretation of Personal Health Messages from Social Media

Nut Limsopatham and Nigel Collier
Department of Theoretical and Applied Linguistics
University of Cambridge
Cambridge, UK
{nl347,nhc30}@cam.ac.uk

ABSTRACT

Recent attempts have been made to utilise social media platforms, such as Twitter, to provide early warning and monitoring of health threats in populations (i.e. Internet biosurveillance). It has been shown in the literature that a system based on keyword matching that exploits social media messages could report flu surveillance well ahead of the Centers of Disease Control and Prevention (CDC). However, we argue that a simple keyword matching may not capture semantic interpretation of social media messages that would enable healthcare experts or machines to extract and leverage medical knowledge from social media messages. In this paper, we motivate and describe a new task that aims to tackle this technology gap by extracting semantic interpretation of medical terms mentioned in social media messages, which are typically written in layman's language. Achieving such a task would enable an automatic integration between the data about direct patient experiences extracted from social media and existing knowledge from clinical databases, which leads to advances in the use of community health experiences in healthcare services.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

Keywords: Internet Biosurveillance, Medical Concept Coding

1. INTRODUCTION

A social media platform, e.g. Twitter[1] and DailyStrength[2], is one of the key elements of a smart city's knowledge infrastructure, from which the communications among citizens could be leveraged in order to enhance the citizens' quality of life [13, 14]. Existing studies have used the Twitter social media platform as a source for retrieving and monitoring real-world events [18, 19, 29]. The insights mined from social media platforms could be leveraged to enhance different aspects of a smart urban environment, such as safety & security, transportation and public health surveillance [10, 13, 29]. For example, Sakaki et al. [29] used Twitter as a *social sensor* for reporting earthquakes in Japan.

Public health surveillance is one aspect of urban informatics that has the potential to enhance the population's quality of life [4, 11]. Indeed, a social media platform is a potentially rich source of *the voice of the patient* data about experience in healthcare (e.g. benefits and side-effects of drugs and treatments for particular diseases) [23]. Leveraging such information from social media would enable the detection of timely health signals from the voice of the patient. For example, flu surveillance using social media messages can detect outbreaks faster than the system of the Centers for Disease Control and Prevention (CDC), which relies on outpatient reports [31]. In this paper, we argue that discussions in social media platforms are useful sources of medical insights about direct experience of patients with healthcare. However, automatically extracting these medical insights from social media messages is challenging, because of the lexical and grammatical variability of the language [3, 15, 23]. Specifically, social media messages are typically abbreviated, ambiguous and informal. In this work, we introduce the task of *medical concept coding for social media messages*, where a medical concept coding system aims to identify the mentions of health information, such as conditions, treatments, medications and behaviours written in layman's language in social media messages, and map them to medical concepts in standard ontologies such as SNOMED-CT [33] and UMLS Metathesaurus[3]. The development of such a system is one of the objectives of our SIPHS project (Semantic Interpretation of Personal Health messages for generating public health summaries) that aims to develop an intelligence system for exploiting the voice of the patient data in medical research.

We envisage that such a medical concept coding system would enable a machine to understand and make inferences on health information (e.g. diseases and treatments) discussed in social media platforms, which would lead to an enhanced quality of population's health. In particular, such a system allows the voice of the patient data to be automatically integrated with existing clinical databases, and hence be used for different applications of Internet bio-surveillance, such as early alerting of adverse drug reactions in populations and infectious disease surveillance [10, 23].

In this paper, we describe the medical concept coding task, the challenges it poses and its relation to current research

[1] http://twitter.com
[2] http://www.dailystrength.org/
[3] http://www.nlm.nih.gov/research/umls/

CIKM'15, October 19–23, 2015, Melbourne, Australia.

Table 1: Examples of the mappings between social media messages and medical concepts.

Social media message	Description of corresponding medical concept
No way I'm gettin any sleep 2nite	Insomnia (SNOMED ID: 193462001)
kept me up for days	Insomnia (SNOMED ID: 193462001)
OMG!! LET ME SLEEP	Insomnia (SNOMED ID: 193462001)
I can't stay focused	Unable to concentrate (SNOMED ID: 60032008)
can't even focus forreal	Unable to concentrate (SNOMED ID: 60032008)
I should be studying for but literally can't	Unable to concentrate (SNOMED ID: 60032008)
DRUG makes u skinny	Weight loss (SNOMED ID: 89362005)
still tired as shit	Fatigue (SNOMED ID: 84229001)
wiggin out a little bit	Fidgeting (SNOMED ID: 247910009)
I'm happiest with _DRUG_	Cheerful mood (SNOMED ID: 112080002)
DRUG made me the most chipper person	Cheerful mood (SNOMED ID: 112080002)
DRUG has me making my roommates bed	Compulsive cleaning (SNOMED ID: 247967000)
DRUG had me literally running into random peoples rooms and OCD cleaning their stuff yesterday afternoon	Compulsive cleaning (SNOMED ID: 247967000)

in information retrieval (IR) and natural language processing (NLP). Indeed, the remainder of this paper is organised as follows. Section 2 defines the medical concept coding task. Section 3 describes the construction of a dataset for evaluating a medical concept coding system. In Section 4, we discuss possible baseline approaches. Finally, we provide concluding remarks in Section 5.

2. TASK DEFINITION

The medical concept coding task aims to enable machines to recognise medical concepts discussed in particular social media messages. Indeed, language understanding by machines requires the ability to recognise when a text refers to a particular medical concept [15]. Given a variable length text, an effective concept coding system should return medical concepts mentioned in the text. For example, a text 'heart disease' may be mapped to medical concept 'Cardiovascular disease' (SNOMED ID: 266894000) when using the SNOMED-CT ontology. In the context of medical records search, recent studies (e.g. [16, 17, 27]) have shown that representing medical records and queries in terms of medical concepts is more effective than a traditional bag-of-words representation.

Although medical concept coding systems exist for clinical and scientific texts (e.g. MetaMap[4][2] and cTAKES[5][30]), medical concept coding for social media messages is a more challenging task because of the unique characteristics of social media messages. Indeed, social media messages are normally short, ambiguous and de-contextualised, while also containing slang and evolving vocabularies [28]. In addition, social media messages are typically posted in layman's language rather than in a formal medical language used in medical literatures and description of medical concepts in medical ontologies (e.g. UMLS and SNOMED-CT) [15].

Table 1 shows examples of the mappings between social media messages and their corresponding medical concepts in the SNOMED-CT ontology, which are annotated by a PhD-level computational linguist. From these examples, we observe the use of non-standard language and abbreviation, such as gettin, 2nite, OMG and OCD. In addition, as shown in Table 1, the lexical similarity between terms in social media messages and in the descriptions of their corresponding medical concepts is rather low. For instance, there is no matched term between a social media message 'No way I'm gettin any sleep 2nite' and the description of its corresponding concept 'Insomnia' (SNOMED ID: 193462001). With these unique characteristics, existing medical concept coding systems may not be effective. Hence, there is a need for the development of a medical concept coding system that is specialised for social media messages.

We model this task as a ranking task, where a medical concept coding system ranks medical concepts based on their similarity with a given social media message. This enables the use of well-established IR evaluation measures for assessing the performance of a medical concept coding system. Initially, we use precision at cut-off 1 (i.e. Precision@1) and Mean Reciprocal Rank (MRR) [6] for evaluation. Precision@1 is a more strict measure, where a system is awarded a score only when it ranks the relevant concept at the top. Meanwhile, MRR is based on a user model where the user wants to see only one relevant concept (i.e. the reciprocal of the rank at which the first relevant concept is viewed in the ranking). For instance, MRR = 0.5 if the first mapped concept is wrong but the second is correct.

3. DATASET CONSTRUCTION

In this section, we discuss how we construct a dataset for developing and evaluating a medical concept coding system for social media messages. In order to construct a large-scale dataset, we use social media messages (i.e. tweets) from the Twitter social media platform, which has been used for several different applications (e.g. real-time identification of earthquakes [29], detection and retrieval of real-world events [18, 19] and pharmacovigilance [23]).

We collect tweets using the Twitter Streaming API[6]. In particular, using the Streaming API, we filter only tweets related to Internet bio-surveillance (as demonstrated by the mentioning of keywords related to health information e.g. diseases and treatments). We highlight two important challenges for filtering tweets. Firstly, different terms can be used to refer to a particular medical concept. Hence infor-

[4]http://metamap.nlm.nih.gov/
[5]http://ctakes.apache.org/
[6]https://dev.twitter.com/streaming/public

mation filtering techniques such as [1] could be adopted to enhance recall. Secondly, assuming that we have a set of possible keywords for the medical concept of interests, variant forms that include incorrect spellings (e.g. influanza instead of influenza) might not be captured. We deal with this by adapting the approach of [26]. In addition, we remove spam from the collected tweets as suggested in [9]. For privacy concerns, the remaining tweets are anonymised by replacing numbers, user IDs, URIs, locations, email addresses, dates and drug names with appropriate tokens, such as *_NUMBER_*, *_DRUG_*. The anonymised tweets are then annotated by at least two biomedical linguists to identify medical concepts and the span of their mentions in the tweets. The annotators are asked to look for mentions of medical concepts related to Internet bio-surveillance, as well as the sentiment of the tweets regarding the identified medical concepts. We use SNOMED-CT as a reference ontology when annotating the tweets. The agreement among the annotators are measured using Cohen's kappa [5]. The development of this dataset is ongoing; however, we plan to make it (including both tweet IDs and their annotations) publicly available in the near future.

There exist datasets for evaluating medical concept coding systems for medical documents (e.g. BioCreative [37], I2B2 [35] and CLEF ShARE [34]). However, they focused on only medical documents, such as clinical records and medical articles, rather than social media messages. Meanwhile, O'Conner et al. [23] introduced a dataset of tweets related to adverse drug reaction (ADR). This ADR dataset contains only 1,008 tweets mentioning medical concepts. In addition, it focuses on adverse drug reaction, which is only one aspect of the Internet bio-surveillance.

4. BASELINE APPROACHES

Next, we discuss possible baseline approaches. Existing studies (e.g. [7, 8, 36]) mostly focused on extracting medical concepts from medical documents. For example, Gobbel et al. [8] proposed a naive Bayesian-based technique to map phrases from clinical notes to medical concepts in the SNOMED-CT ontology. Wang et al. [36] identified medical concepts regarding adverse drug events in electronic medical records. However, it was shown that existing NLP techniques performed poorly when dealing with Twitter messages, because of compressed nature of tweets [28].

O'Connor et al. [23] investigated the extraction of medical terms from Twitter messages. In particular, they proposed to use the Lucene retrieval engine[7] to retrieve medical concepts that could be potentially mapped to a given Twitter message, when mapping between Twitter messages and medical concepts. Nevertheless, as previously shown in Table 1, the lexicon of medical concepts mentioned in tweets is likely to be different from that of the description of medical concepts in a standard ontology. Therefore, such an approach may not be effective.

Recent advances in the development of techniques for learning high-quality word vector representations (i.e. distributed word representations), such as continuous bags of words (CBOW)[8] [21] and global vectors (GloVe)[9] [25], enable a machine to capture semantic similarity between words. Indeed, these distributed word representations have been effectively applied in different systems that achieve state-of-the-art performances for several NLP tasks, such as sentiment analysis [32], machine translation [20] and named entity recognition [24]. Therefore, assuming that distributed word representations could bridge the semantic similarity gap between terms used in layman description of medical concepts in social media messages and terms used in the formal description of medical concepts in a medical ontology, a possible effective baseline is to rank medical concepts based on the similarity between the learned distributed word representations of a given tweet and the description of each medical concept.

Another research direction for tackling the medical concept coding task is to use statistical machine translation (MT) techniques. Phrase-based MT models (e.g. [12, 22]) have been shown to be effective in translation between languages, as they learn local term dependencies, such as collocations, re-orderings, insertions and deletions. Koehn et al. [12] showed that a phrase-based MT technique markedly outperformed traditional word-based MT techniques on several benchmarks. A medical concept coding system may use a phrase-based MT technique to translate from *Twitter language* (e.g. 'No way I'm gettin any sleep 2nite') to *formal medical language* (e.g. 'insomnia'), when mapping a tweet to a medical concept.

5. CONCLUSIONS

We have introduced a medical concept coding task that aimed to create semantic interpretation of social media messages. In particular, the task is to map health information in social media messages to medical concepts in a medical ontology. Achieving this task will enable a machine to understand and reason about health of population from the voice of the patient, and hence lead to their improved health quality.

We also showed that, however, medical concept coding for social media messages was challenging, due to the unique characteristics of social media messages, which were typically short, de-contextualised, ambiguous, and often contained slang, as well as evolving vocabularies. In addition, we detailed our construction of a large-scale dataset using social media messages collected from Twitter. Moreover, we discussed possible baseline approaches that could be applied to tackle this coding task, and the challenges to be overcome, which could stimulate further research.

Acknowledgements

The authors gratefully acknowledge funding from the EPSRC (grant number EP/M005089/1).

6. REFERENCES

[1] Allan, James. Incremental relevance feedback for information filtering. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 1996.

[2] A. R. Aronson and F.-M. Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[3] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang. How noisy social media text, how diffrnt social media sources. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013.

[7] `http://lucene.apache.org/`

[8] `https://code.google.com/p/word2vec/`

[9] `http://nlp.stanford.edu/projects/glove/`

[4] M. A. Barrett, O. Humblet, R. A. Hiatt, and N. E. Adler. Big data and disease prevention: From quantified self to quantified communities. *Big data*, 1(3), 168-175, 2013.

[5] J. Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

[6] N. Craswell. Mean reciprocal rank. In *Encyclopedia of Database Systems*, pages 1703–1703. Springer, 2009.

[7] P. L. Elkin, D. A. Froehling, D. L. Wahner-Roedler, S. H. Brown, and K. R. Bailey. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. *Annals of Internal Medicine*, 156(1_Part_1):11–18, 2012.

[8] G. T. Gobbel, R. Reeves, S. Jayaramaraja, D. Giuse, T. Speroff, S. H. Brown, P. L. Elkin, and M. E. Matheny. Development and evaluation of raptat: a machine learning system for concept mapping of phrases from medical narratives. *Journal of biomedical informatics*, 48:54–65, 2014.

[9] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: The underground on 140 characters or less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, 2010.

[10] D. M. Hartley, N. P. Nelson, R. Arthur, P. Barboza, N. Collier, N. Lightfoot, J. Linge, E. Goot, A. Mawudeku, L. Madoff, L. Vaillant, R. Walters, R. Yangarber, J. Mantero, C. D. Corley, and J. S. Brownstein. An overview of internet biosurveillance. *Clinical Microbiology and Infection*, 19(11):1006–1013, 2013.

[11] A. Hipp, D. Adlakha, R. Gernes, A. Kargol, and R. Pless. Learning from Outdoor Webcams: Surveillance of Physical Activity Across Environments. In *Proceedings of the NSF Workshops on Big Data and Urban Informatics*. University of Illinois at Chicago, 2014.

[12] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics, 2003.

[13] H. Kukka, V. Kostakos, T. Ojala, J. Ylipulli, T. Suopajärvi, M. Jurmu, and S. Hosio. This is not classified: Everyday information seeking and encountering in smart urban spaces. *Personal Ubiquitous Comput.*, 17(1):15–27, 2013.

[14] N. Limsopatham, M. Albakour, C. Macdonald, and I. Ounis. Tweeting behaviour during train disruptions within a city. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, 2015.

[15] N. Limsopatham, and N. Collier. Adapting Phrase-based Machine Translation to Normalise Medical Terms in Social Media Messages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2015.

[16] N. Limsopatham, C. Macdonald, and I. Ounis. A Task-Specific Query and Document Representation for Medical Records Search. In *Advances in Information Retrieval*, 2013.

[17] N. Limsopatham, C. Macdonald, and I. Ounis. Inferring Conceptual Relationships to Improve Medical Records Search. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, 2013.

[18] R. McCreadie, C. Macdonald, I. Ounis, M. Osborne, and S. Petrovic. Scalable distributed event detection for twitter. In *Proceedings of the IEEE International Conference on Big Data*, 2013.

[19] D. Metzler, C. Cai, and E. Hovy. Structured event retrieval over microblog archives. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012.

[20] T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[22] F. J. Och and H. Ney. The alignment template approach to statistical machine translation. *Computational linguistics*, 30(4):417–449, 2004.

[23] K. O'Connor, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. L. Smith, and G. Gonzalez. Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA Annual Symposium Proceedings*, 2014.

[24] A. Passos, V. Kumar, and A. McCallum. Lexicon infused phrase embeddings for named entity resolution. *arXiv preprint arXiv:1404.5367*, 2014.

[25] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014.

[26] P. Pimpalkhute, A. Patki, A. Nikfarjam, and G. Gonzalez. Phonetic spelling filter for keyword selection in drug mention mining from social media. In *Proceedings of AMIA Summits on Translational Science*, 2014.

[27] Y. Qi and P.-F. Laquerre. Retrieving Medical Records with "sennamed": NEC Labs America. In *Proceedings of the 21st Text REtrieval Conference*, 2012.

[28] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

[29] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

[30] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. 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, 2010.

[31] C. W. Schmidt. Trending now: using social media to predict and track disease outbreaks. *Environ Health Perspect*, 120(1):30–33, 2012.

[32] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

[33] K. A. Spackman, K. E. Campbell, and R. A. Côté. Snomed rt: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, 1997.

[34] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. Jones, et al. Overview of the share/clef ehealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, 2013.

[35] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[36] X. Wang, G. Hripcsak, M. Markatou, and C. Friedman. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, 16(3):328–337, 2009.

[37] A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman. Biocreative task 1a: gene mention finding evaluation. *BMC bioinformatics*, 6(Suppl 1):S2, 2005.