Adapting Phrase-based Machine Translation to Normalise Medical Terms in Social Media Messages

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

Abstract

Previous studies have shown that health reports in social media, such as DailyStrength and Twitter, have potential for monitoring health conditions (e.g. adverse drug reactions, infectious diseases) in particular communities. However, in order for a machine to understand and make inferences on these health conditions, the ability to recognise when laymen's terms refer to a particular medical concept (i.e. text normalisation) is required. To achieve this, we propose to adapt an existing phrase-based machine translation (MT) technique and a vector representation of words to map between a social media phrase and a medical concept. We evaluate our proposed approach using a collection of phrases from tweets related to adverse drug reactions. Our experimental results show that the combination of a phrase-based MT technique and the similarity between word vector representations outperforms the baselines that apply only either of them by up to 55%.

1 Introduction

Social media, such as DailyStrength[1] and Twitter[2], is a fast growing and potentially rich source of *voice of the patient* data about experience in terms of benefits and side-effects of drugs and treatments (O'Connor et al., 2014). However, natural language understanding from social media messages is a difficult task because of the lexical and grammatical variability of the language (Baldwin et al., 2013; O'Connor et al., 2014). Indeed, language understanding by machines requires the ability to recognise when a phrase refers to a particular concept. Given a variable length phrase, an effective system should return a concept with the most similar meaning. Table 1 shows examples of mappings between Twitter phrases and medical concepts. For example, a Twitter phrase 'No way I'm gettin any sleep 2nite' might be mapped to the medical concept 'Insomnia' (SNOMED:193462001), when using the SNOMED-CT ontology (Spackman et al., 1997). The success of the mapping between social media phrases and formal medical concepts would enable an automatic integration between patient experiences and biomedical databases (Limsopatham and Collier, 2015). We refer to this mapping from social media phrases to medical concepts as *medical term normalisation*, which aims to determine the unique identifier of a medical concept that is mentioned in different forms in a free-text (Morgan et al., 2008).

Existing works, e.g. (Elkin et al., 2012; Gobbel et al., 2014; Wang et al., 2009), mostly focused on identifying medical concepts in medical documents. For example, Gobbel et al. (2014) proposed a naïve Bayesian-based technique to map phrases from clinical notes to medical concepts in the SNOMED-CT ontology. Wang et al. (2009) identified medical concepts regarding adverse drug events in electronic medical records. On the other hand, O'Connor et al. (2014) investigated the normalisation of medical terms in Twitter messages. In particular, they proposed to use the Lucene retrieval engine[3] to retrieve medical concepts that could be potentially mapped to a given Twitter phrase, when mapping between Twitter phrases and medical concepts.

In contrast, we argue that the medical text normalisation task can be achieved by using well-established phrase-based MT techniques, where we translate a text written in *a social media language* (e.g. 'No way I'm gettin any sleep 2nite') to a text written in *a formal medical language* (e.g.

[1] http://www.dailystrength.org/
[2] http://twitter.com
[3] http://lucene.apache.org/

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1675–1680,
Lisbon, Portugal, 17-21 September 2015.

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)
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)

'Insomnia') and then calculate the similarity between the translated phrase and the description of a medical concept. Indeed, in this work we investigate an effective adaptation of phrase-based MT to map a Twitter phrase to a medical concept. Moreover, we propose to combine the adapted phrase-based MT technique and the similarity between word vector representations to effectively map a Twitter phrase to a medical concept.

The main contributions of this paper are threefold:

1. We investigate the adaptation of phrase-based MT to map a Twitter phrase to a SNOMED-CT concept.
2. We propose to combine our adaptation of phrase-based MT and the similarity between word vector representations to map Twitter phrases to formal medical concepts.
3. We thoroughly evaluate the proposed approach using phrases from our collection of tweets related to the topic of adverse drug reactions (ADRs).

2 Related Work

Phrase-based MT models, e.g. (Koehn et al., 2003; Och and Ney, 2004), have been shown to be effective in translation between languages, as they learn local term dependencies, such as collocations, reorderings, insertions and deletions. Koehn et al. (2003) showed that a phrase-based MT technique markedly outperformed traditional word-based MT techniques on several benchmarks. In this work, we adapt the phrase-based MT technique of Koehn et al. (2003) for the medical text normalisation task. In particular, we use the phrase-based MT technique to translate phrases from *Twitter language* to *formal medical language*, before mapping the translated phrases to medical concepts based on the ranked similarity of their word vector representations.

Traditional approaches for creating word vector representations treated words as atomic units (Mikolov et al., 2013b; Turian et al., 2010). For instance, the one-hot representation used a vector with a length of the size of the vocabulary, where one dimension is on, to represent a particular word (Turian et al., 2010). Recently, techniques for learning high-quality word vector representations (i.e. distributed word representations) that could capture the semantic similarity between words, such as continuous bags of words (CBOW) (Mikolov et al., 2013b) and global vectors (GloVe) (Pennington et al., 2014), have been proposed. 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 MT (Mikolov et al., 2013a) and named entity recognition (Passos et al., 2014). In this work, beside using word vector representations to measure the similarity between translated Twitter phrases and the description of medical concepts, we use the similarity between word vector representations of the original Twitter phrase and the description of a medical concept to augment the adapted phrase-based MT technique.

3 Medical Term Normalisation

We discuss our adaptation of phrase-based MT for medical text normalisation in Section 3.1. Section 3.2 introduces our proposed approach for combining similarity score of word vector representations with the adapted phrase-based MT technique.

3.1 Adapting Phrase-based MT

We aim to learn a translation between a Twitter phrase (i.e. a phrase from a Twitter message) and a formal medical phrase (i.e. the description of a medical concept). For a given Twitter phrase phr_t, we find a suitable medical phrase phr_m using *a translation score*, based on a phrase-based model, as follows:

$$score_{translation}(phr_m|phr_t) = p(phr_m|phr_t) \quad (1)$$

where $p(phr_m|phr_t)$ can be calculated using any phrase-based MT technique, e.g. (Koehn et al., 2003; Och and Ney, 2004). We then rank translated phrases phr_m based on this translation score. The top-k translated phrases are used for identifying the corresponding medical concept.

However, the translated phrase phr_m may not be exactly matched with the description of any target medical concepts. We propose two techniques to deal with this problem. For the first technique, we rank the target concepts based on the cosine similarity between the vector representation of phr_m and the vector representation of the description of each concept $desc_c$:

$$sim_{cos}(phr_m, desc_c) = \frac{V_{phr_m} \cdot V_{desc_c}}{||V_{phr_m}|| \times ||V_{desc_c}||} \quad (2)$$

where V_{phr_m} and V_{desc_c} are the vector representations of phr_m and $desc_c$, respectively. Any technique for creating word vector representations (e.g. one-hot, CBOW and GloVe) can be used. Note that if a phrase (e.g. phr_m) contains several terms, we create a vector representation by summing the value of the same dimension of the vector representation of each word (i.e. element-wise addition).

On the other hand, the second technique also incorporates the ranked position r of the translated phrase phr_m when translated from the original phrase phr_t using Equation (1). Indeed, the second technique calculates the similarity score as follows:

$$sim_{rcos}(phr_m, desc_c) = \frac{1}{r} \cdot \frac{V_{phr_m} \cdot V_{desc_c}}{||V_{phr_m}|| \times ||V_{desc_c}||} \quad (3)$$

3.2 Combining Similarity Score with Phrase-based MT

As discussed in Section 2, word vector representations (e.g. created by CBOW or GloVe) can capture semantic similarity between words by itself. Hence, we propose to map a Twitter phrase phr_t to a medical concept c, which is represented with a description $desc_c$, by linearly combining the cosine similarity, between vector representations of the Twitter phrase phr_t and the description $desc_c$, with the similarity score computed using one of the adapted phrased-based MT techniques (introduced in Section 3.1), as follows:

$$sim_{combine}(phr_t, desc_c) = \frac{V_{phr_t} \cdot V_{desc_c}}{||V_{phr_t}|| \times ||V_{desc_c}||} + MT_a(phr_t, desc_c) \quad (4)$$

where $MT_a(phr_t, desc_c)$ is calculated using one of the adapted phrase-based MT techniques described in Section 3.1.

4 Experimental Setup

4.1 Test Collection[4]

To evaluate our approach, we use a collection of 25 million tweets related to adverse drug reactions (ADRs), from cognitive enhancers (Hanson et al., 2013) and anti-depressants (Schneeweiss et al., 2010). These tweets were collected using the Twitter Streaming API[5] by filtering on the name of a particular set of drugs that can have adverse reactions to the patients. Note that terms regarding adverse drug reaction (e.g. insomnia) were not used for capturing tweets. From this collection, we use 201 ADR phrases and their corresponding SNOMED-CT concepts annotated by a PhD-level computational linguist. These phrases were anonymised by replacing numbers, user IDs, URIs, locations, email addresses, dates and drug names with appropriate tokens e.g. *_NUMBER_*.

4.2 Evaluation Approach

We conduct experiments using 10-fold cross validation, where the Twitter phrases are randomly divided into 10 separated folds. We address this task as a ranking task, where we aim to rank the medical concept with the highest similarity score, e.g. calculated using Equation (2), at the top rank. Hence, we evaluate our approach using Mean Reciprocal Rank (MRR) measure (Craswell, 2009), which is an information retrieval measure based on the user model where the user wants to see

[4]The gold-standard mapping between the Twitter phrases and the SNOMED-CT concepts are available on Zenodo.org (DOI: http://dx.doi.org/10.5281/zenodo.27354)

[5]https://dev.twitter.com/streaming/public

only one relevant concept. In particular, MRR is based on the the reciprocal of the rank at which the first relevant concept is viewed in the ranking (e.g. MRR = 0.5 if the first mapped concept is wrong but the second is correct). We limit our evaluation at top 5 of the ranking (i.e. MRR-5). In addition, we compare the significant difference between the performance achieved by our proposed approach and the baselines using the paired t-test ($p < 0.05$).

4.3 Word Vector Representation

We use three different techniques, including one-hot, CBOW and GloVe, to create word vector representations used in our approach (see Section 3). In particular, the vocabulary for creating the one-hot representation includes all terms in the Twitter phrases and the descriptions of the target SNOMED-CT concepts. Meanwhile, we create word vector representations based on CBOW and GloVe by using the word2vec[6] and GloVe[7] implementations. We learn the vector representations from the collections of tweets and medical articles, respectively, using window size of 10 words. The tweet collection (denoted *Twitter*) contains 419,702,147 English tweets, which are related to 11 drug names and 6 cities, while the medical article collection (denoted *BMC*) includes all medical articles from the BioMed Central[8]. For both CBOW and GloVe, we create vector representations with vector sizes 50 and 200, respectively.

4.4 Learning Phrase-based Model

We use the phrase-based MT technique of Koehn et al. (2003), as implemented in the Moses toolkit (Koehn et al., 2007)[9] with default settings, to learn to translate from the Twitter language to the medical language. In particular, when training the translator, we show the learner pairs of the Twitter phrases and descriptions of the corresponding SNOMED-CT concepts.

5 Experimental Results

We evaluate 6 different instantiations of the proposed approach discussed in Section 3, including:

1. *bestMT*: set $k = 1$, when finding the translated phrase phr_m for a Twitter phrase phr_t (Equation (1)), before ranking target medical concepts for the translated phrase phr_m using Equation (2).
2. *top5MT*: similar to *bestMT*, but set $k = 5$.
3. *top5MTr*: similar to *top5MT*, but also consider the rank position of the translate phrases when ranking the target medical concepts by using Equation (3).
4. *bestMT+vSim*: incorporate with the ranking generated from *bestMT*, the cosine similarity between the vector representations of the Twitter phrase phr_t and the description $desc_c$ of target medical concepts by using Equation (4).
5. *top5MT+vSim*: similar to *bestMT+vSim*, but use the ranking from *top5MT*.
6. *top5MTr+vSim*: similar to *bestMT+vSim*, but use the ranking from *top5MTr*.

Another baseline is *vSim*, where we consider only the cosine similarity between the vector representations of the Twitter phrase phr_t and the description $desc_c$ of target medical concepts.

Table 2 compares the performance of these 6 instantiations and the *vSim* baseline in terms of MRR-5. We firstly observe that for the *vSim* baseline, excepting for word vector representation with vector size 50 learned using GloVe from the Twitter collection, word vector representations learned using either CBOW or GloVe are more effective than the one-hot representation. However, the difference between the MRR-5 performance is not statistically significant ($p > 0.05$, paired t-test). In addition, word vector representations learned either using CBOW or GloVe with vector size 200 is more effective than those with vector size 50.

Next, we find that our adaptation of phrase-based MT (i.e. *bestMT*, *top5MT* and *top5MTr*) significantly ($p < 0.05$) outperforms the *vSim* baseline. For example, with the one-hot representation, *top5MT* (MRR-5 0.2491) and *top5MTr* (MRR-5 0.2458) perform significantly ($p < 0.05$) better than *vSim* (MRR-5 0.1675) by up to 49%. Meanwhile, when using word vector representations with the vector size 200 learned using GloVe from the BMC collection, *top5MT* (MRR-5 0.2638) significantly ($p < 0.05$) outperforms *vSim* with either the GloVe vector representation (MRR-5 0.1869) or the one-hot representation (MRR-5 0.1675). We observe the similar trends in performance

[6] https://code.google.com/p/word2vec/

[7] http://nlp.stanford.edu/projects/glove/

[8] http://www.biomedcentral.com/about/datamining

[9] http://www.statmt.org/moses/

Table 2: MRR-5 performance of the proposed approach and the baselines. Significant differences ($p < 0.05$) compared to the cosine similarity (*vSim*) baselines with the one-hot representation, and with the corresponding distributed word representation (e.g. CBOW or GloVe) are denoted $^{\triangle}$ and $^{\blacktriangle}$, respectively.

Approach	One-hot	BMC				Twitter			
		CBOW		GloVe		CBOW		GloVe	
		50	200	50	200	50	200	50	200
vSim	0.1675	0.1771	0.1896	0.1840	0.1869	0.1812	0.1813	0.0936	0.1807
bestMT	0.2232	0.1926	0.2070	0.1803	0.2500$^{\triangle}$	0.2014	0.2047	0.1258	0.2138
top5MT	0.2491$^{\triangle}$	**0.1994**	0.2104	0.1879	**0.2638**$^{\triangle\blacktriangle}$	**0.2037**	0.2095	**0.1322**	0.2362
top5MTr	0.2458$^{\triangle}$	0.1982	0.2109	**0.1894**	0.2617$^{\triangle}$	**0.2037**	**0.2096**	**0.1322**	0.2310
bestMT+vSim	0.2420$^{\triangle}$	0.1910	0.1953	0.1860	0.2532$^{\triangle}$	0.1891	0.1954	0.1078	0.2374
top5MT+vSim	0.2556$^{\triangle}$	0.1916	**0.2144**	0.1726	0.2600$^{\triangle}$	0.1978	0.2068	0.1079	0.2405$^{\triangle}$
top5MTr+vSim	**0.2594**$^{\triangle}$	0.1861	0.2070	0.1802	0.2590$^{\triangle}$	0.1959	0.2027	0.1129	**0.2406**$^{\triangle}$

when using vector representations learned from the Twitter collection. These results show that our adapted phrase-based MT techniques are effective for the medical term normalisation task.

In addition, we observe the effectiveness of our combined approach (i.e. *bestMT+vSim*, *top5MT+vSim* and *top5MTr+vSim*), as it further improves the performance of the adapted phrase-based MT (i.e. *bestMT*, *top5MT* and *top5MTr*, respectively), when using the one-hot representation. For example, *top5MTr+vSim* achieves the MRR-5 of 0.2594, while the MRR-5 of *top5MTr* is 0.2458. However, the performance difference is not statistically significant. Meanwhile, when using the CBOW and GloVe vectors, the achieved performance is varied based on the collection (i.e. BMC or Twitter) used for learning the vectors and the size of the vectors.

6 Conclusions

We have introduced our approach that adapts a phrase-based MT technique to normalise medical terms in Twitter messages. We evaluate our proposed approach using a collection of phrases from tweets related to ADRs. Our experimental results show that the proposed approach significantly outperforms an effective baseline by up to 55%. For future work, we aim to investigate the modelling of learned vector representation, such as CBOW and GloVe, within a phrase-based MT model when normalising medical terms.

Acknowledgements

The authors gratefully acknowledge Nestor Alvaro (Sokendai, Japan) for providing access to the Twitter/SNOMED-CT annotations which were used to derive the test collection used in these experiments. The derived dictionary and a representative sample of the word vector representations (CBOW and GloVe at 200d) are made available on Zenodo.org (DOI: http://dx.doi.org/10.5281/zenodo.27354). We wish to thank funding support from the EPSRC (grant number EP/M005089/1).

References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

Nick Craswell. 2009. Mean reciprocal rank. In *Encyclopedia of Database Systems*, pages 1703–1703. Springer.

Peter L Elkin, David A Froehling, Dietlind L Wahner-Roedler, Steven H Brown, and Kent R Bailey. 2012. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. *Annals of Internal Medicine*, 156(1_Part_1):11–18.

Glenn T Gobbel, Ruth Reeves, Shrimalini Jayaramaraja, Dario Giuse, Theodore Speroff, Steven H Brown, Peter L Elkin, and Michael E Matheny. 2014. Development and evaluation of raptat: a machine learning system for concept mapping of phrases from medical narratives. *Journal of biomedical informatics*, 48:54–65.

Carl L Hanson, Scott H Burton, Christophe Giraud-Carrier, Josh H West, Michael D Barnes, and Bret Hansen. 2013. Tweaking and tweeting: exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students. *Journal of medical Internet research*, 15(4).

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. 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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Nut Limsopatham and Nigel Collier. 2015. Towards the semantic interpretation of personal health messages from social media. *Proceedings of the 1st Workshop on Understanding the City with Urban Informatics (UCUI 2015) in conjunction with CIKM 2015* (in press). `https://www.repository.cam.ac.uk/handle/1810/249275`

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Alexander A Morgan, Zhiyong Lu, Xinglong Wang, Aaron M Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, et al. 2008. Overview of biocreative ii gene normalization. *Genome biology*, 9(Suppl 2):S3.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational linguistics*, 30(4):417–449.

Karen O'Connor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, and Graciela Gonzalez. 2014. Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA Annual Symposium Proceedings*, volume 2014, page 924. American Medical Informatics Association.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. *arXiv preprint arXiv:1404.5367*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543.

Sebastian Schneeweiss, Amanda R Patrick, Daniel H Solomon, Colin R Dormuth, Matt Miller, Jyotsna Mehta, Jennifer C Lee, and Philip S Wang. 2010. Comparative safety of antidepressant agents for children and adolescents regarding suicidal acts. *Pediatrics*, pages peds–2009.

Kent A Spackman, Keith E Campbell, and Roger A Côté. 1997. Snomed rt: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Xiaoyan Wang, George Hripcsak, Marianthi Markatou, and Carol Friedman. 2009. 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.