Lexical normalization of user-generated medical forum data

Anne Dirkson
Leiden University
a.r.dirkson@liacs.leidenuniv.nl

Suzan Verberne
Leiden University
s.verberne@liacs.leidenuniv.nl

Gerard van Oortmerssen
Leiden University
g.van.oortmerssen@liacs.leidenuniv.nl

Wessel Kraaij
Leiden University
w.kraaij@liacs.leidenuniv.nl

ABSTRACT

In the medical domain, user-generated social media text is increasingly used as a valuable complementary knowledge source to scientific medical literature: it contains the unprompted experiences of the patient. Yet, lexical normalization of such data has not been addressed properly. This paper presents a sequential, unsupervised pipeline for automatic lexical normalization of domain-specific abbreviations and spelling mistakes. This pipeline led to an absolute reduction of out-of-vocabulary terms of 0.82% and 0.78% in two cancer-related forums. Our approach mainly targeted, and thus corrected, medical concepts. Consequently, our pipeline may significantly improve downstream IR tasks.

CCS CONCEPTS

• **Computing methodologies → Information extraction**; • **Applied computing → Consumer health**; **Health informatics**;

KEYWORDS

lexical normalization, social media, patient forum, domain-specific

ACM Reference Format:
Anne Dirkson, Suzan Verberne, Gerard van Oortmerssen, and Wessel Kraaij. 2018. Lexical normalization of user-generated medical forum data. In *Proceedings of Dutch-Belgian Information Retrieval Workshop (DIR2018).* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

In recent years, user-generated data from social media have been used extensively for medical text mining and information retrieval (IR) [4]. This user-generated data encapsulates a vast amount of knowledge, which has been used for a range of health-related applications, such as the tracking of public health trends [13] and the detection of adverse drug responses [12]. However, the extraction of this knowledge is complicated by non-standard and colloquial language use, typographical errors, phonetic substitutions, and misspellings [3, 11]. Social media text is generally noisy, and the complex medical domain aggravates this challenge [4]. The unique domain-specific terminology on forums cannot be captured by professional clinical terminologies because laypersons and healthcare professionals express health-related concepts differently [16].

Despite these challenges, normalization is one of the least explored topics in social media health language processing [4]. Medical lexical normalization methods, i.e. abbreviation expansion [6] and spelling correction [5, 10], have mostly been developed for clinical records or notes, as these also contain an abundance of domain-specific abbreviations and misspellings. However, social media text presents distinct challenges [4, 11] and cannot be tackled with these methods.

At the ACL W-NUT workshop in 2015, the best performing system for lexical normalization of generic social media combined rule-based and learning-based techniques [14]. Recently, Sarker [11] developed a modular pipeline that outperformed this system. His pipeline includes a customizable back-end module for domain-specific normalization, which employs spelling correction specifically for medical terms. However, it does not take into account that specialized forums often contain highly specific terms which may be excluded from the vocabulary. These terms are often essential for the task at hand (e.g. a novel drug name) and should thus not be 'corrected'. Additionally, Sarker [11] did not tackle domain-specific abbreviation expansion.

Thus, to further improve the quality of medical forum data, in this paper we will present two sequential domain-specific modules for lexical normalization of user-generated data, targeting abbreviations and spelling mistakes. The aim of this paper is two-fold. Firstly, we investigate to what extent these lexical normalization techniques can improve the quality of the patient forum text. Secondly, we apply these techniques to the second patient forum to test to what extent they are generalizable to other cancer-related medical forums.

2 DATA

2.1 Medical forum data

The first forum is a Facebook community, moderated by GIST Support International, an international patient forum for patients with Gastrointestinal Stromal Tumor (GIST). The data was collected in 2015 in collaboration with TNO. The second forum is the sub-reddit community on cancer, dating from 16/09/2009 until 02/07/2018.[1] It was scraped using the Pushshift Reddit API.[2] The data was collected in batches by looping over the timestamps in the data.

2.2 Abbreviations lexicon

Abbreviations were manually extracted from 500 randomly selected posts from the GIST data. This resulted in 47 unique abbreviations. For each abbreviation, two annotators firstly individually determined the correct expansion term, with an absolute agreement of 85.4%. Hereafter, they agreed on the correct form together. If

DIR2018, Nov 2018, Leiden, the Netherlands
2018.

[1] www.reddit.com/r/cancer
[2] https://github.com/pushshift/api

	# Tokens	# Posts	Median length of post (IQR)
GIST forum	1,225,741	36,722	20 (35)
Reddit forum	4,520,074	274,532	11 (18)

Table 1: Raw data. The number of tokens and the median length of a post were calculated without punctuation.

ambiguous or context-dependent, the abbreviation was removed. For this reason, five abbreviations were removed.

2.3 Annotated data for spelling correction

The same 500 randomly selected posts were split into two sets of 250 posts: a tuning and a test set for detecting spelling mistakes. Each token was classified as a mistake (1) or not (0) by the first author. A second annotator checked if any of the mistakes were false positives. The first subset contained 34 unique non-word errors, equal to 0.39% of the tokens. Real-word errors, valid words used in the incorrect context, were not included. For the test set, these 34 mistakes and a tenfold of randomly selected correct words (340) with the same word length distribution were selected. The second subset contained 23 unique mistakes, equal to 0.31% of the tokens in the set. The tuning set consisted of these 23 mistakes combined with a tenfold of randomly selected correct words (230) with the same word length distribution. The tuning set was split in a stratified manner into 10 folds for cross-validation.

Combined, the two sets contained 55 unique mistakes: two mistakes occurred in both sets. The corrections of these mistakes were annotated individually by two annotators and then agreed on together. The absolute agreement was 89.0%. 8 mistakes were removed due to ambiguity (e.g. 'annonse' or 'gon'), resulting in 47 unique mistakes for evaluating the spelling correction algorithms.

3 METHODS

3.1 Preprocessing

To protect the privacy of users, in-text personal pronouns have been replaced as much as possible using a combination of the NLTK names corpus and part-of-speech tags (NNP and NNPS). Additionally, URLs and email addresses were replaced by the strings -url- and -email- using regular expressions. Furthermore, text was lowercased and tokenized using NLTK. The first modules of the normalization pipeline of Sarker [11] were employed: converting British to American English spelling and the lexicon-based normalization of generic abbreviations. Some forum-specific additions were made: Gleevec (British variant: Glivec) was included in the first step and one generic abbreviation expansion that clashed with a domain-specific expansion was removed (i.e. 'temp' defined as *temperature* instead of *temporary*). Moreover, the Sarker dictionary was lowercased and tokenized prior to preprocessing.

3.2 Abbreviation expansion

A simple lexicon lookup was used to expand the abbreviations in the data.

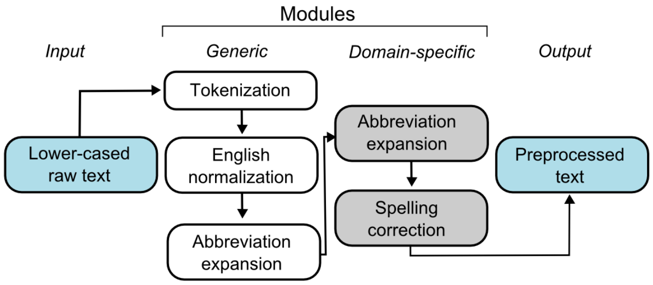

Figure 1: Sequential processing pipeline

3.3 Spelling correction

We used the method by Sarker [11] (S1) as a baseline for spelling correction. His method combines normalized absolute Levenshtein distance (NAE) with Metaphone phonetic similarity and language model similarity. For the latter, distributed word representations (skip-gram word2vec) of three large Twitter datasets were used. It was compared with absolute Levenshtein distance (NAE), normalized as was done in S1, and relative Levenshtein distance (RE). Both were also explored with a penalty (-1) for differing first letters. Additionally, we investigated a version of Sarker's algorithm without language model similarity (S2).

We manually constructed a decision process, inspired by the work by Beeksma [1], for detecting spelling mistakes. The decision process makes use of a token's frequency in the corpus, and the similarity with possible replacements. The underlying idea is that if a word is common within the domain-specific language or there is no similar enough candidate available, it is unlikely to be a mistake.

To ensure generalisability, we opted for an unsupervised, data-driven method that does not rely on the construction of a specialized vocabulary. For measuring similarity and correcting terms, the generic CELEX lexicon [2] was combined with all corpus tokens surpassing the frequency threshold. The latter are considered only after the CELEX terms and in order of frequency (from high to low). Of the candidates with the highest similarity score, the first is selected.

To optimize the decision process, a 10-fold cross validation grid search of the maximum relative corpus frequency [1E-6, 2.5E-6, 5E-6, 1E-5, 2E-5, 4E-5] and maximum relative edit distance (0.15 to 0.25 with 0.01 increments) was conducted with the tuning set. The choice of grid was based on previous work by Walasek [15] and Beeksma [1]. The loss function used to tune the parameters was the $F_{0.5}$ score, which places more weight on precision than the F_1 score. We believe it is more important to not alter correct terms, than to retrieve incorrect ones.

3.4 Evaluating data quality

The percentage of out-of-vocabulary (OOV) terms is used as an estimation of the quality of the data: less OOV-terms and thus more in-vocabulary (IV) terms reflects cleaner data. To calculate the number of OOV terms, a merged vocabulary was created by combining the standard English lexicon CELEX [2], the NCI Dictionary of Cancer Terms [7], the generic and commercial drug names from

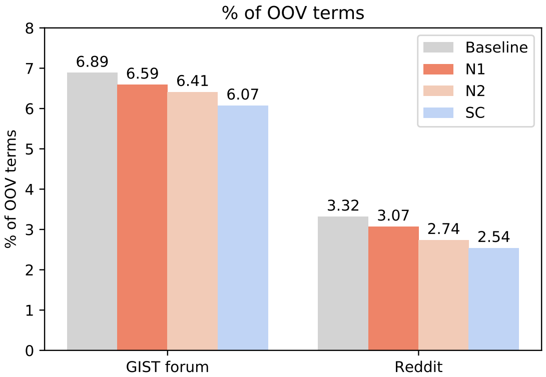

Figure 2: Number of OOV-terms with sequential modules. N1: Generic abbreviation expansion [11]. N2: Domain-specific abbreviation expansion. SC: Spelling correction.

the RxNorm [8], the ADR lexicon used by Nikfarjam et al. [9] and our abbreviation expansions. [3]

4 RESULTS

4.1 Abbreviation expansion

The baseline % of OOV-terms was higher for the GIST data (6.9%) than the Reddit data (3.3%). The most effective reduction of OOV-terms in both forums was achieved by combined generic and domain-specific abbreviation expansion (N1+N2) (see Fig 2). This was slightly more effective in the Reddit (-0.58%) than the GIST data (-0.47%) (see Fig. 2).

The additional domain-specific abbreviation expansion replaced 4747 terms distributed over 3756 posts (18.7% of the data) in the GIST forum and 18688 terms in 16479 posts (6.0% of the data) in the Reddit forum. The associated OOV-term reduction was 0.18% and 0.33% resp. The replacements did not appear concentrated in a small number of posts in either forum: respectively 81.3% and 88.9% of the posts with replacements had a single replacement.

31 of the 36 abbreviations found in the GIST forum were also present in the Reddit forum, indicating that these abbreviations are to some extent generalizable between cancer-related forums. The abbreviations that were not present in the cancer sub-reddit were: hpfs (*high power fields*), vit (*vitamin*), gf (*girlfriend*), mg/d (*mg/day*) and til (*until*). There was also large overlap (80%) between the ten most common abbreviation expansions in the forums. For the Reddit forum, chemotherapy (69.9%) was by far the most common expansion. Although a common treatment for many cancers, it is an uncommon treatment for GIST, which explains the relative low frequency (5.7%) for the GIST forum.

4.2 Spelling correction

Detecting spelling mistakes. The grid search resulted in a max. corpus frequency of 5E-06 and a max. similarity score of 0.19 (see Table 2). This combination attained the maximum $F_{0.5}$ score for all folds. Despite a low recall on the test set (0.38), the precision was 1. Thus, although mistakes may be missed, no correct terms are falsely marked as errors. Unfortunately, this does mean that some common mistakes, like oncogolgist, are missed (see Table 3).

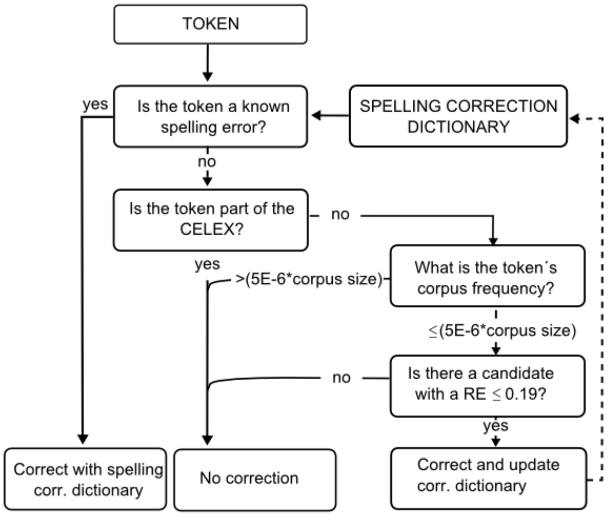
TOKEN
Is the token a known spelling error?
SPELLING CORRECTION DICTIONARY
yes
no
Is the token part of the CELEX?
no
yes
>(5E-6*corpus size)
What is the token´s corpus frequency?
≤(5E-6*corpus size)
Is there a candidate with a RE ≤ 0.19?
no
yes
Correct with spelling corr. dictionary
No correction
Correct and update corr. dictionary

Figure 3: Decision process for spelling corrections. RE: Relative Edit Distance

		Recall	Precision	F_1	$F_{0.5}$	AUC
CELEX	Test	0.94	0.51	0.66	0.56	0.92
Decision process	Validation	0.62	0.76	0.67	0.72	0.80
	Test	0.38	1.0	0.55	0.75	0.69

Table 2: Detection of spelling mistakes. The average of a 10-fold CV was taken for the validation set.

False negatives	abdomin	oncogolgist	metastisis	thanx
True positives	oncolgy	clenical	metastized	surgry

Table 3: Examples of false negatives (i.e. missed mistakes) and true positives (i.e. found mistakes) found in the test set using mistake detection with the decision process

	NAE	NAE+P	RE	RE+P	S1	S2
Accuracy	59.6%	59.6%	**66.0%**	**66.0%**	23.4%	19.1%
Duration (s)	6.09	7.29	3.84	4.07	257.00	237.42

Table 4: Spelling correction. NAE: normalized absolute edit distance. +P: with first-letter penalty. RE: relative edit distance. S1: Sarker's algorithm S2: S1 without language model similarity. Duration was measured over an average of 5 runs.

Comparing spelling correction algorithms. Relative edit distance (RE) was the most accurate spelling correction algorithm (66.0%) (see Table 4). The first-letter penalty did not improve the accuracy.

[3] available at urlhttps://github.com/AnneDirkson/lex_normalization

Mistake	gleevac	opnion	sutant	kontrol
Correction	gleevec	opinion	sutent	control
NAE	**gleevec**	option	mutant	**control**
NAE+P	**gleevec**	option	**sutent**	kowtow
RE	**gleevec**	**opinion**	mutant	**control**
RE+P	**gleevec**	**opinion**	**sutent**	kestrel
S1	colonic	option	mutant	contr
S2	gleeful	option	mutant	controls

Table 5: Examples of spelling correction results. NAE: normalized absolute edit distance. +P: with first-letter penalty. RE: relative edit distance. S1: Sarker's algorithm. S2: S1 without the language model.

Since the corrections of four mistakes did not occur in the vocabulary, the upper bound of accuracy was 91.5%. Interestingly, the two versions of Sarker's method (S1 and S2) managed to correct only 23.4% and 19.1% of the mistakes respectively. This showcases the limitations of using generic social media normalization techniques in the medical domain.

Evaluating the spelling correction module. In the GIST data, 3367 mistakes were replaced with 2601 unique terms. The mistakes often concern important medical terms. The ten most frequent corrections were: gleevec (17x), oncologist (13x), diagnosed (10x), positive (8x), stivarga (8x), imatinib (8x), metastasized (7x), regorafenib (7x) and tumors (7x). Gleevec, stivarga, imatinib and regorafenib are cancer medications.

In the Reddit forum, 5238 mistakes were replaced with 4161 unique terms, of which the most prevalent were: metastasized (10x), treatment (10x), diagnosed (10x), adenocarcinoma (10x), symptoms (9x), immunotherapy (9x), lymphoma (8x), patients (8x), dexamethasone (8x) and cannabinoids (8x). Thus, our module appears to effectively target medical terms.

The reduction in OOV-terms was higher for the GIST (0.34%) than for the Reddit forum (0.20%) (See Fig. 2). Furthermore, our method only targets infrequent spelling mistakes: in both forums, all corrected spelling mistakes occurred only once.

5 DISCUSSION

For domain-specific abbreviation expansion and sequential spelling correction, the combined reduction in OOV-terms was 0.59% and 0.54% for the GIST and Reddit forum resp. Although this reduction may seem minor, our approach mainly targets medical concepts, which are highly relevant for downstream tasks such as named entity extraction. The pipeline appears generalizable for cancer-related forums: it resulted in comparable reductions in OOV-terms for both forums.

The generic lexical normalization pipeline by Sarker [11] does not appear to suffice for normalizing health-related user-generated text. We identified 36 additional domain-specific abbreviations in our data that were not corrected in their method. Moreover, our analysis revealed that their spelling correction algorithm performed poorly compared to both relative and absolute Levenshtein distance. One must note, however, that the test set excluded real-word errors, slang and ambiguous errors.

Our study has a number of limitations. Firstly, the use of OOV-terms as a proxy for quality of the data relies heavily on the vocabulary that is chosen and, moreover, does not allow for differentiation between correct and incorrect substitution of words. In the future, we will instead opt for extrinsic performance measures to investigate the utility of our approach. Secondly, our data-driven spelling correction could lead to the ‘correction’ of spelling mistakes with other spelling mistakes. This possibility cannot be excluded entirely, but is countered by sorting the corpus tokens on frequency. A larger tuning set could perhaps improve the thresholding.

6 CONCLUSION

Our sequential unsupervised pipeline can improve the quality of text data from medical forum posts. Future work will explore the impact of our pipeline on task performance using established benchmark data from diverse medical forums.

REFERENCES

[1] M. Beeksma. 2017. *Computer: how long have I got left?* Master's thesis. Radboud University, Nijmegen, the Netherlands.

[2] G. Burnage, R.H Baayen, R. Piepenbrock, and H. van Rijn. 1990. CELEX: A Guide for Users. (1990).

[3] E. Clark and K. Araki. 2011. Text Normalization in Social Media: Progress, Problems and Applications for a Pre-processing System of Casual English. *Procedia Soc Behav Sci* 27 (2011), 2–11. https://doi.org/10.1016/j.sbspro.2011.10.577

[4] G. Gonzalez-Hernandez, A. Sarker, K. O 'Connor, and G. Savova. 2017. Capturing the Patient's Perspective: a Review of Advances in Natural Language Processing of Health-Related Text. *Yearbook of medical informatics* (2017), 214–217. https://doi.org/10.15265/IY-2017-029

[5] K.H. Lai, M. Topaz, F.R. Goss, and L. Zhou. 2015. Automated misspelling detection and correction in clinical free-text records. (2015). https://doi.org/10.1016/j.jbi.2015.04.008

[6] D.L. Mowery, B.R. South, L. Christensen, J. Leng, L.M. Peltonen, S. Salanterä, H. Suominen, D. Martinez, S. Velupillai, N. Elhadad, G. Savova, S.r Pradhan, and W. W. Chapman. 2016. Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth Challenge 2013, Task 2. *Journal of Biomedical Semantics* (2016). https://doi.org/10.1186/s13326-016-0084-y

[7] National Cancer Institute. [n. d.]. NCI Dictionary of Cancer Terms. https://www.cancer.gov/publications/dictionaries/cancer-terms

[8] National Library of Medicine (US). [n. d.]. RxNorm. https://www.nlm.nih.gov/research/umls/rxnorm/

[9] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association : JAMIA* 22, 3 (2015), 671–81. https://doi.org/10.1093/jamia/ocu041

[10] J. Patrick, M. Sabbagh, S. Jain, and H. Zheng. 2010. Spelling correction in clinical notes with emphasis on first suggestion accuracy. *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining* (2010), 2–8.

[11] A. Sarker. 2017. A customizable pipeline for social media text normalization. *Social Network Analysis and Mining* 7, 45 (2017). https://doi.org/10.1007/s13278-017-0464-z

[12] A. Sarker, R. Ginn, A. Nikfarjam, K. O‘Connor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics* 54 (2015), 202–212. https://doi.org/10.1016/J.JBI.2015.02.004

[13] A. Sarker, K. O‘Connor, R. Ginn, M. Scotch, K. Smith, D. Malone, and G. Gonzalez. 2016. Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. *Drug Safety* 39, 3 (2016), 231–240. https://doi.org/10.1007/s40264-015-0379-4

[14] D. Supranovich and V. Patsepnia. 2015. IHS_RD: Lexical Normalization for English Tweets. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*. 78–81.

[15] N. Walasek. 2016. *Medical Entity Extraction on Dutch forum data in the absence of labeled training data.* Master's thesis. Radboud University, Nijmegen, the Netherlands.

[16] Q. Zeng and T. Tse. 2006. Exploring and developing consuming health vocabulary. *J Am Med Inform Assoc* 13, 1 (2006), 24–29. https://doi.org/10.1197/jamia.M1761.A