Int. j. inf. tecnol. (August 2022) 14(5):2417–2422
https://doi.org/10.1007/s41870-022-00995-w

ORIGINAL RESEARCH

Building a reverse dictionary with specific application to the COVID-19 pandemic

Bushra Siddique[1] · **M. M. Sufyan Beg**[1]

Received: 2 April 2022 / Accepted: 2 May 2022 / Published online: 10 June 2022

Abstract A Reverse Dictionary maps a natural language description to corresponding semantically appropriate words. It is of assistance, particularly to the language producers, in finding the correct word for a concept in mind while writing/speaking. As the COVID-19 pandemic intensely impacted almost all the functionalities across the globe, texts on this subject appear in a significant amount in various forms, including news updates, awareness and safety articles, notices and circulars, research articles, social media posts, etc. A Reverse Dictionary on this subject is a requisite in view of the following reasons, hence addressed. Firstly, the varied text forms involve a diverse range of language producers ranging from professional doctors to the general mass. Secondly, the COVID-19 pandemic's glossary is more specific than the general English language, hence unfamiliar to the language producers. We have carried out an implementation based on the Wordster Reverse Dictionary architecture, owing to its outperformance of the commercial Onelook Reverse Dictionary benchmark. We report an accuracy of 0.49 based on top-3 system responses. To address the limitations of the current implementation, we bring into consideration Zadeh's paradigm of the Computational Theory of Perceptions. Notably, the compilation of the COVID-19 glossary as a part of this study is another contribution in view that it is of assistance to the concerned readers.

Keywords Reverse Dictionary · RD · COVID-19 · Coronavirus · Wordster RD · Information retrieval

✉ Bushra Siddique
bushrasiddique006@gmail.com

[1] Department of Computer Engineering, Aligarh Muslim University, Aligarh, India

1 Introduction

A forward dictionary contains mappings of a word to its meaning. It serves the purpose of looking up an unknown word while reading a text, thus addressing the needs specific to readers. The needs of language producers (speakers and writers) are, however, different. While composing a piece of text, the problem often faced is not to look up the meaning of a word but to get/recall an appropriate word corresponding to a meaningful phrase. Given its organization, a forward dictionary falls short in addressing this problem, thereby introducing the concept of a Reverse Dictionary (may be henceforth referred to as RD). A Reverse Dictionary maps a phrase to semantically appropriate words. Specifically, it addresses the Tip-of-Tongue (TOT) problem [3], meaning the word is on the tongue tip but the person is not able to articulate it. Several works exist in the literature addressing the problem of Reverse Dictionary (could be referred to in [24])and can be grouped into the following categories: Information Retrieval (IR) System based Approach [2, 5, 8, 22, 23], Graph Based Approach [7, 21, 27], Mental Dictionary based Approach [34, 35], Vector Space Model based Semantic Analysis Approach [4, 12], Neural Language Model based Approach [1, 9, 10, 14, 15, 18, 19, 29, 32] .

A few commercial RD applications also exist; namely, Onelook.com [17] and reversedictioanry.org [20], out of which the latter is the most popular and serves as the benchmark for assessing the quality of the research-based RD works. Amongst the Information Retrieval (IR) System-based RD works, Wordster RD [22] is reported to outperform Onelook.com. We have built a specific Reverse Dictionary on the basis of the architecture of the Wordster RD.

In this study, we have build a RD solution in the specific context of the coronavirus pandemic. As this outbreak was recognized as a global health emergency, many texts addressing it started flooding. A study reports that as many as 23,634 unique published articles were indexed on Web of Science and Scopus in just the initial phase of the pandemic.[1] The text included varied forms, the primary being the news updates. Others include articles on awareness, diagnostic symptoms, safety measures, medical treatments etc. With due course of time, as this pandemic affected almost all the functionalities across the globe, the form of the relevant text became far diverse including modified government regulations, education system guidelines, notices and circulars in various organizations, research articles etc. Correspondingly, the set of language producers for these texts encompassed a variety of people ranging from doctors, journalists, political leaders, researchers, office staff, medical staff, social workers, non-government organizations, patients, education system staff, and the general mass on social platforms.

Owing to its specificity, the glossary of the coronavirus pandemic is unfamiliar to the language producers. Hence, the problem of missing an apt technical term while writing on this subject is inevitable. This problem becomes exaggerated given the variety of language producers including the professionals as well as the non-professionals. Consequently, a Reverse Dictionary on this subject is a requisite, and we address the same through this paper. In the process, we have compiled a database of the COVID-19 glossary from various Internet sources.[2,3, 4,5,6,7] This could be accounted as another contribution of this paper, as it could assist the interested readers on this subject.

The work carried out in this study is remarkably different from the similar-appearing task of medical terminology mapping [11, 16, 28]. Firstly, the former deals with finding an apt term in semantic coherence with a natural language input description, whereas, the latter deals with the identification and extraction of medical terms in natural language text. Secondly, in terms of objective, the former assists the language producers in general whereas the latter assists the medical professionals in particular. Lastly, in terms of scope, the former is associated with medical (such as *Severe Acute Respiratory Infection (SARI)* and *Acute Respiratory Stress Syndrome (ARDS)*)terms as well as non-medical terms (such as *lockdown* and *Work From Home(WFH)*) related to the COVID-19 pandemic whereas the latter is associated with medical terms only.

The paper is organized as follows. In Sect. 2, we brief the architecture of the COVID-19 Reverse Dictionary, outlining the key algorithms. In Sect. 3, we provide the details of the implementation and results. Finally, the paper is concluded.

2 COVID-19 reverse dictionary: architecture and working

The architecture of the COVID-19 RD is based on the Wordster RD [22] which in turn is based on the Information Retrieval System's architecture. Given a forward dictionary consisting of pairs of word and its definition, the Wordster RD accepts a user input description in natural language, converts it into a Boolean expression query, extracts definitions based on a semantic similarity measure and then outputs the words corresponding to the definitions in the form of a ranked list. Accordingly, these constitute the primary steps in the COVID-19 RD implementation. The flow of the system is illustrated in Fig. 1. Along the same lines, a Reverse Dictionary with specific application to English idioms is reported in [25]. We employ the primary algorithms of the Wordster RD in our implementation. The modules of the COVID-19 RD is outlined as under:

1. Building the Reverse Map Set: The Reverse Map Set (RMS) applies to the terms in the vocabulary of a given forward dictionary. Specifically, a forward dictionary consists of word-definitions pairs and the set of terms appearing in the word definitions constitute its vocabulary.
   For a term t, $RMS(t)$ consists of the vocabulary words whose definition contains the term t. Mathematically,
   $$RMS(t) = \{W_1, W_2, \ldots, W_n\}\ s.t.\ t \in def(W_i), i = 1, 2, \ldots, n \quad (1)$$
   For example, consider the word *numismatics* meaning *the study or collection of coins, banknotes, and medals*. The term set of the definition after stemming and stop words removal consist of the terms *study, collect, coin, banknote, medal*. Accordingly, *numismatics* appears in the *RMS* of the these terms.
2. Query Generation: For a given user input description, a query in the form of a Boolean expression is generated. Initially, the query is formulated as an ANDed expression of the terms appearing in the user input

[1] https://www.natureindex.com/news-blog/how-coronavirus-is-changing-research-practices-and-publishing.

[2] https://uvahealth.com/services/covid19-glossary.

[3] https://www.kff.org/glossary/covid-19-outbreak-glossary/.

[4] https://www.cedars-sinai.org/blog/covid-19-vocabulary.html.

[5] https://www.englishclub.com/vocabulary/coronavirus-covid19.php.

[6] https://www.thehindu.com/sci-tech/health/the-hindu-explains-what-are-some-of-the-key-terms-being-used-to-describe-the-novel-coronavirus-outbreak/article31768617.ece.

[7] https://www.tmc.edu/news/2020/05/covid-19-crisis-catalog-a-glossary-of-terms/.

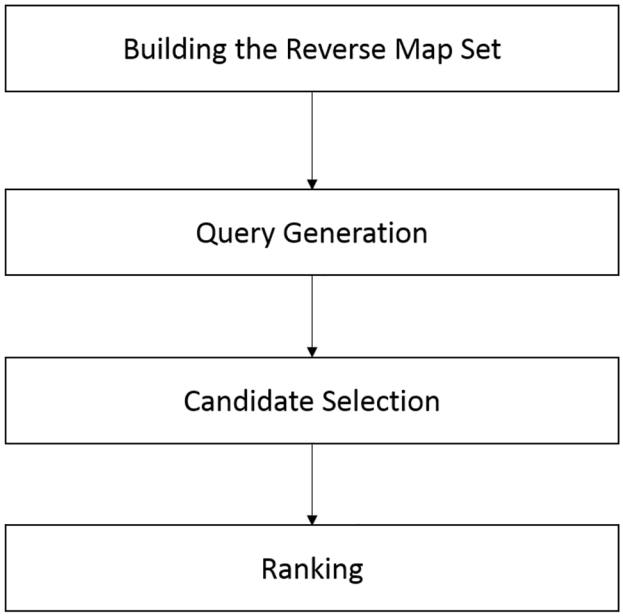

Fig. 1 A high level representation of COVID-19 Reverse Dictionary

description (after it is processed for basic natural language processing tasks).

Let $t_1, t_2, \ldots, t_n$ be the terms appearing in the processed user input description U. Correspondingly, the query Q is represented as,

$$Q = t_1 \odot t_2 \cdots \odot t_n \tag{2}$$

such that $\odot$ correspond to the *AND* operator. In case the required number of potential candidates are not fetched, the query is expanded using lexical relatedness like synonymy. The lexically related terms are connected to the query terms via the OR operator. Let, the synonyms of the query term t_i be $t_{i_{s1}}, t_{i_{s2}}, \ldots, t_{i_{sk}}$. The expanded query Q' takes the form as shown below:

$$Q' = (t_{1_{s1}} \oplus t_{1_{s2}} \oplus \cdots \oplus t_{1_{sm_1}}) \odot \cdots (t_{n_{s1}} \oplus t_{n_{s2}} \oplus \cdots \oplus t_{n_{smn}}) \tag{3}$$

such that $\oplus$ correspond to the *OR* operation.

3. Candidate Selection: For a given Boolean expression query, the candidate selection module extracts candidate definitions for ranking. This is done by executing the set operations on the RMS of the query terms (refer Eq. 1).

   The set operation corresponding to the $\odot$ and $\oplus$ operators in the query expression is . (*intersection* operation) and + (denoted by *union* operation) respectively. Performing these operations on the Reverse Map Set of the query terms result in a set of those words whose definition contain **all** the query terms (either in original from or connected through lexical relatedness), thus qualify to be relevant to the user input.

   Correspondingly, the definition phrases of these vocabulary words form the set of potential candidates. Assume the following to be a sample query,

$$Q = (t_1 \oplus t_2 \oplus t_3) \odot (t_4 \oplus t_5) \odot (t_6) \tag{4}$$

   Then, the potential candidates correspond to the set of following words:

$$[RMS(t_1) + RMS(t_2) + RMS(t_3)].[RMS(t_4) + RMS(t_5)].[RMS(t_6)] \tag{5}$$

4. Ranking: The ranking module ranks the given set of potential candidates on the basis of the similarity score of the candidate definition with the query. The similarity score consists of following two components:

   – Term Similarity, t_s: depicts similarity between each pair of terms of the definition (potential candidate) and the query. In the Wordster RD, the employed measure is Wu and Palmer simiarity [33] according to which the similarity t_s between terms a and b is calculated as under:

$$t_s = \frac{2 * D(LCA(a,b))}{(D(a) + E(b))} \tag{6}$$

   such that $D(t)$ is depth of the term t in the WordNet [13] hierarchy and $LCA(a, b)$ is the least common ancestor of the terms a and b.

   – Term Importance, t_i: depicts the importance of a term in a phrase (definition/query). This is based upon the structure of the parse tree of the phrase. The term importance value for the term t appearing in the phrase p is calculated as follows:

$$t_i(t,p) = \frac{(D(Z(p)) - D(t))}{D(Z(p))} \tag{7}$$

such that $Z(p)$ is the parse tree of the phrase p, $D(a)$ gives the depth of the entity a.

The two components, t_s and t_i are aggregated as S_w for the term t_D of the definition, D, and the term $t_Q b$ of the query, Q using the following equation:

$$S_w(t_D, D, t_Q, Q) = t_i(t_D, D) * t_i(t_Q, Q) * t_s(t_D, t_Q) \tag{8}$$

where the terms a and b appear in the candidate definition D and the user input U respectively. The weighted similarity score S_w for each pair of terms is then used to calculate the overall similarity using [6]. This final score forms the basis of ranking the candidates in the order of decreasing values. The result set consists of the vocabulary words corresponding to the candidate definitions in the ranked list.

3 Implementation details and evaluation results

We have compiled a COVID-19 dictionary dataset consisting of 110 words and their definitions from relevant links on the Internet.[8,9,10,11,12,13] While compiling, if more than one definitions for a word is available from multiple web links, it is taken into account. Specifically, of the total of 110 words, 37 words have multiple definitions.

To prepare the test set, we have randomly selected 30 words from the compiled data set. These words are given to a set of 4 potential users of varied backgrounds, medical as well as non-medical. Each user is asked to write a phrasal description for each test word. The word's actual definition is provided to the user only if it is required by him/her for clarification. The collected descriptions are checked for quality by providing them to the user other than the one who wrote it. If the test word is not correctly guessed based on the test description, rewriting is done. In this way, 120 test descriptions are collected, 4 for each test word.

3.1 Comparison benchmarks

As reported in the previous RD works, we have considered the commercial Onelook.com Reverse Dictionary [17] for comparison. Also, we have taken into consideration the recently reported WantWords Online RD [19]. As mentioned (in the Introduction section), the vocabulary addressed by the COVID RD implementation encompasses terms related to the pandemic. Thus, it includes medical-related terms (such as Severe Acute Respiratory Infection (SARI) etc.) as well as non-medical related terms (such as lockdown etc.).

We concluded that while the considered benchmarks could generate related responses for non-medical test terms definitions, the generated responses for the medical-related test term definitions were highly unrelatable. This in view that these RDs are for general English vocabulary and do not cater to specific applications. The overall accuracy of the test set for both is too low to consider. Onelook.com nevertheless achieved higher accuracy than WantWords. In view of this, we carried out a manual-based evaluation of our implementation as reported in the following section.

[8] https://uvahealth.com/services/covid19-glossary.

[9] https://www.kff.org/glossary/covid-19-outbreak-glossary/.

[10] https://www.cedars-sinai.org/blog/covid-19-vocabulary.html.

[11] https://www.englishclub.com/vocabulary/coronavirus-covid19.php.

[12] https://www.thehindu.com/sci-tech/health/the-hindu-explains-what-are-some-of-the-key-terms-being-used-to-describe-the-novel-coronavirus-outbreak/article31768617.ece.

[13] https://www.tmc.edu/news/2020/05/covid-19-crisis-catalog-a-glossary-of-terms/.

Table 1 Sample successful system responses

S. no.	Test word	Test description	Rank
1	Viral	Caused because of virus	2
2	Vax	Term for vaccine	1
3	Epidemic	Fast spreading of any disease over a particular geographical area	3
4	Person under investigation (PUI)	Person who is carefully observed before testing	1
5	Spanish flu	Disease caused by H1N1 influenza virus	1

Table 2 Sample unsuccessful system responses

S. no.	Test word	Test description
1	Pre-symptomatic	Person who is not showing symptoms of disease
2	Fomite	Non living carriers of an infecting agent
3	PCR Test	Test performed to detect presence of genetic material from a covid virus
4	Lockdown	Method for isolating peoples by shutting down activities
5	Presumptive positive case	Person tested positive by private hospital but not by government

3.2 Manual-based evaluation results

For the prepared test set, we have obtained an accuracy of 0.48 based on the top-3 responses generated by the system. Given the number of participants, the reported accuracy value is the average over four runs. This implies that for about half of test user descriptions, the implementation can generate the sought word within the top-3 positions. Tables 1 and 2 lists a few samples of the successful and unsuccessful system responses, respectively.

4 Limitations and future directions

The current implementation takes into account the lexical semantics of the phrases: user input descriptions as well as the dictionary descriptions. The phrases are treated as a bag-of-keywords, hence, the implicit semantics is disregarded. For example, consider the test description at S. No. 5 of Table 2. The terms 'but' and 'not' in the test description "Person tested positive by private hospital but not by government" conveys the specific intent of the user. If the description is treated as a bunch of keywords, such user intents could not be addressed.

To address this gap, we need to base our RD solution on a paradigm capable of dealing with implicit semantics in natural language. In view of this, we bring into consideration the paradigm of Computational Theory of Perceptions (CTP) [30] proposed by L.A. Zadeh in which the objects of computation are words rather than numbers. Based on this paradigm, we propose to incorporate the concept of Precisiated Natural Language (PNL) [31] in building a Reverse Dictionary solution. A specific instance of the same is reported in our study [26].

5 Conclusion

As the COVID-19 pandemic has affected almost all the functionality across the globe, text in varied forms are found to appear in significant amount. The diversity of text forms encompasses a variety of language producers. Unlike the general English language vocabulary, the COVID-19 glossary is different and unfamiliar to the language producers and hence the problem of finding an appropriate technical term during composition of text/speech is inescapable. In view of this, we implement a Reverse Dictionary on the subject of COVID-19 pandemic. As a part of the implementation, we compile a data set of COVID-19 glossary which provides assistance to the readers as well. The implementation carried out is based on the framework of Wordster Reverse Dictionary and an accuracy of 0.48 is reported based on top-3 responses generated by the implementation. The scope of improvement in the task of RD lies in incorporating paradigms capable of dealing with implicit semantics, particularly Zadeh's Computational Theory of Perceptions (CTP).

Funding No funding was received.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Agrawal A, Shanly KA, Vaishnaw K, Singh M (2021) Reverse dictionary using an improved cbow model. In: 8th ACM IKDD CODS and 26th COMAD, 420
2. Bilac S, Watanabe W, Hashimoto T, Tokunaga T, Tanaka H (2004) Dictionary search based on the target word description. In: Proceedings of the tenth annual meeting of the association for NLP (NLP2004), pp 556–559
3. Brown R, McNeill D (1966) The"tip of the tongue'' phenomenon. J Verbal Learn Verbal Behav 5(4):):325–337
4. Calvo H, Méndez O, Moreno-Armendáriz MA (2016) Integrated concept blending with vector space models. Comput Speech Lang 40:79–96
5. Crawford HV, Crawford J (1997) Reverse electronic dictionary using synonyms to expand search capabilities. US Patent 5,649,221
6. Dao TN, Simpson T (2005) Measuring similarity between sentences. The Code Project
7. Dutoit D, Nugues P (2002) A lexical database and an algorithm to find words from definitions. In: ECAI, pp 450–454
8. El-Kahlout ID, Oflazer K (2004) Use of wordnet for retrieving words from their meanings. In: Proceedings of the global Wordnet conference (GWC2004), pp 118–123
9. Hedderich MA, Yates A, Klakow D, De Melo G (2019) Using multi-sense vector embeddings for reverse dictionaries. arXiv preprint arXiv:1904.01451
10. Hill F, Cho K, Korhonen A, Bengio Y (2015) Learning to understand phrases by embedding the dictionary. Trans Assoc Comput Linguist 4:17–30
11. Lauría EJ, March AD (2011) Combining Bayesian text classification and shrinkage to automate healthcare coding: a data quality analysis. J Data Inf Qual (JDIQ) 2(3):1–22
12. Méndez O, Calvo H, Moreno-Armendáriz MA (2013) A reverse dictionary based on semantic analysis using wordnet. Mexican international conference on artificial intelligence. Springer, New York, pp 275–285
13. Miller GA (1995) Wordnet: a lexical database for English. Commun ACM 38(11):39–41
14. Morinaga Y, Yamaguchi K (2018) Improvement of reverse dictionary by tuning word vectors and category inference. International conference on information and software technologies. Springer, New York, pp 533–545
15. Morinaga Y, Yamaguchi K (2020) Improvement of neural reverse dictionary by using cascade forward neural network. J Inf Process 28:715–723
16. Nie L, Zhao YL, Akbari M, Shen J, Chua TS (2014) Bridging the vocabulary gap between health seekers and healthcare knowledge. IEEE Trans Knowl Data Eng 27(2):396–409
17. Onelook.com: OneLook Reverse Dictionary. http://www.onelook.com/
18. Pilehvar MT (2019) On the importance of distinguishing word meaning representations: a case study on reverse dictionary mapping. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1, (Long and Short Papers), pp 2151–2156
19. Qi F, Zhang L, Yang Y, Liu Z, Sun M (2020) Wantwords: an open-source online reverse dictionary system. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp 175–181
20. Reversedictionary.org: Reverse dictionary. http://reversedictionary.org/
21. Reyes-Magaña J, Bel-Enguix G, Sierra G, Gómez-Adorno H (2019) Designing an electronic reverse dictionary based on two word association norms of English language. In: Proceedings of electronic lexicography in the 21st century conference, pp 865–880
22. Shaw R, Datta A, VanderMeer D, Dutta K (2013) Building a scalable database-driven reverse dictionary. IEEE Trans Knowl Data Eng 25(3):528–540
23. Shete PS, Patil G (2018) Intelligent reverse dictionary based on clustering. In: 2018 fourth international conference on computing communication control and automation (ICCUBEA), pp 1–9. IEEE
24. Siddique B, Beg MMS (2018) A review of reverse dictionary: finding words from concept description. International conference

on next generation computing technologies. Springer, New York, pp 128–139
25. Siddique B, Beg MS (2021) i-RD: a reverse dictionary for English Idioms. In: 2021 10th international conference on internet of everything, microwave engineering, communication and networks (IEMECON), pp 1–5. IEEE
26. Siddique B, Beg MS (2022) Adjective phrases in PNL and its application to reverse dictionary. IEEE Access
27. Thorat S, Choudhari V (2016) Implementing a reverse dictionary, based on word definitions, using a node-graph architecture. arXiv preprint arXiv:1606.00025
28. Wang Y, Patrick J (2008) Mapping clinical notes to medical terminologies at point of care. In: Proceedings of the workshop on current trends in biomedical natural language processing, pp 102–103
29. Yan H, Li X, Qiu X (2020) Bert for monolingual and cross-lingual reverse dictionary. arXiv preprint arXiv:2009.14790
30. Zadeh LA (1999) From computing with numbers to computing with words. from manipulation of measurements to manipulation of perceptions. IEEE Trans Circuits Syst I Fundam Theory Appl 46(1):105–119
31. Zadeh LA (2004) Precisiated natural language (pnl). AI Mag 25(3):74
32. Zheng L, Qi F, Liu Z, Wang Y, Liu Q, Sun M (2020) Multi-channel reverse dictionary model. Proc AAAI Conf Artif Intell 34:312–319
33. Zhibiao W, Palmer M (1994) Verb semantics and lexical selection. In: Proceedings of the 32nd annual meeting on association for computational linguistics, pp 133–138
34. Zock M, Bilac S (2004) Word lookup on the basis of associations: from an idea to a roadmap. In: Proceedings of the workshop on enhancing and using electronic dictionaries, pp 29–35. Association for Computational Linguistics
35. Zock M, Schwab D (2008) Lexical access based on underspecified input. In: Proceedings of the workshop on cognitive aspects of the Lexicon, pp 9–17. Association for Computational Linguistics