

RESEARCH

Open Access



Combining string and phonetic similarity matching to identify misspelt names of drugs in medical records written in Portuguese

Hegler Tissot^{1*} and Richard Dobson^{1,2,3}

From UK Healthcare Text Analysis Conference (HealTAC 2018)
Manchester, UK. 18–19 April 2018

Abstract

Background: There is an increasing amount of unstructured medical data that can be analysed for different purposes. However, information extraction from free text data may be particularly inefficient in the presence of spelling errors. Existing approaches use string similarity methods to search for valid words within a text, coupled with a supporting dictionary. However, they are not rich enough to encode both typing and phonetic misspellings.

Results: Experimental results showed a joint string and language-dependent phonetic similarity is more accurate than traditional string distance metrics when identifying misspelt names of drugs in a set of medical records written in Portuguese.

Conclusion: We present a hybrid approach to efficiently perform similarity match that overcomes the loss of information inherit from using either exact match search or string based similarity search methods.

Keywords: Phonetic similarity, Similarity search, Misspelt names of drugs

Background

There is a large amount of unstructured data being produced by different kinds of information systems, in a variety of formats, due to the advancement of communication and information technologies [1, 2]. Within the clinical domain, Electronic Health Record (EHR) systems are becoming widely adopted, from which information describing the patient's health conditions is often presented and stored in the form of free text notes [3]. Existing text-mining methods aim to extract detailed structured information from clinical narratives, such as drug prescriptions, their variability, and adverse drug reactions [4, 5]. However, free-text is susceptible to typing and phonetic misspellings. Spelling errors of generic drug names can occur in up to one out of six entries

in electronic drug information systems. Such errors are likely to be responsible for up to 12% of adverse drug events, mainly caused by errors during transcription of prescriptions, illegible prescriptions, or drug name confusion [6]. Due to such frequency and the relevance of drug information in clinical tasks, spelling correction becomes crucial to support health care professionals with spelling error-tolerant engine systems.

Similarity comparison algorithms can be used to identify and extract concepts from free text [7] when text is loaded with misspellings. String similarity metrics (e.g. Edit Distance [8] and Jaro-Winkler Distance [9]) can measure similarity between two strings. These functions can be used to compare the elements from the input data source against an existing dictionary in order to identify a possible valid word matching a misspelling. However, existing string similarity algorithms may be inefficient to analyse text loaded with spelling errors because they may

*Correspondence: htissot@ucl.ac.uk

¹Institute of Health Informatics, University College London, London, UK
Full list of author information is available at the end of the article



not necessarily handle specific aspects, such as phonetic errors [10]. In these cases, it is necessary to use phonetic similarity metrics.

In order to overcome the possible loss of information by using exact match search methods to find mentions of drugs within patient records, we propose a hybrid solution coupling string and phonetic similarity metrics to identify misspelt names of drugs. This approach was used to produce a dictionary of misspelt variations. A Trie-based fast similarity search algorithm was then able to identify a broader range of potential candidates as misspelt variations for each drug name.

Use-case

Since July 2013 the Brazilian government tries to address the shortage of doctors, especially in the inner cities and the outskirts of large cities in Brazil, through the hiring of doctors from other countries. With the addition of doctors with distinct language background in the public health system (especially from South and Central America where people originally speak Spanish), a larger number of spelling errors have been found in electronic record systems. Such errors occur mainly due to the similarity of the Portuguese language with other Latin languages (such as Spanish and Italian) [11].

InfoSaude (InfoHealth) [12] is an information system created to manage and track medical records, such as exams, vaccinations, and drug prescriptions. The system is used to meet the needs of 75 public health centres in the city of Florianopolis/Brazil. It integrates different information structures used by the Brazilian Ministry of Health, such as the Outpatient Information System (CIS) and the International Code of Diseases (ICD). The system also generates information for Ambulatory Care Individual Report (RAAI), summarizing data on the type of care, pregnancies, procedures performed on the patient, applied vaccines and drug prescriptions. Whilst maintaining a series of structured information, the system also contains textual fields that are filled by health professionals during patient care.

Although *InfoSaude* has structured information about drug prescriptions, a deeper analysis on drug usage, abuse, or checking whether patients are correctly and effectively making use of the prescribed drugs, relies on the observations registered by the clinicians using free text. However, the textual content of the medical records does not go through any kind of review. Thus, it is common to find a number of spelling and phonetic errors that could harm any further analysis. An information extraction system is expected to overcome this problem in order to avoid information loss.

Approximate string match

The existing similarity match methods range from using basic string similarity distance metrics, which measure

inverse similarity between two text strings by providing an algorithm-specific numerical indication of distance, to the use of more sophisticated methods coupled with the phonetic representation of words in a given language.

Edit Distance (ED) (or Levenshtein Distance) [8] is the most widely known string metric. ED operates between two input strings – $ED(w_1, w_2)$ – and returns the minimum number of operations (single-character edits) required to transform string w_1 into w_2 . Other examples and variations of string similarity metrics include Jaro-Winkler Distance [9], Hamming Distance [13], and *StringSim* [14]. However, string distance measures tend to ignore the relative likelihood errors.

Phonetic representations encode words based on the sound of each letter to translate a string into a canonical form. Soundex [15] is an example of a phonetic matching scheme initially designed for English that uses codes based on the sound of each letter to translate a string into a canonical form of at most four characters, preserving the first letter. In addition, phonetic similarity metrics are able to assign a high score even though comparing dissimilar pairs of strings that produce similar sounds [14, 16]. As the result, phonetically similar entries will have the same (or similar) keys and they can be indexed for efficient search using some hashing method. However, phonetics is language-dependent [17, 18] and solutions for this sort of problems must be specially designed for each specific language.

In addition, fast similarity search approaches have been proposed in order to match free text against large dictionaries or databases, being supported by either indexed database structures [14, 19, 20] or Trie-based (prefix index) approximate matching [21–23]. In an initial experiment, Fuzzy Keyword Search [22] has proved to be efficient by combining Trie-based search with string similarity functions. However, processing time grows exponentially as long as the Edit Distance threshold increases, becoming inefficient for $ED > 2$, which we were able to confirm by comparing the processing time (in milliseconds) spent to perform 1000 searches over a dictionary of 80,000 entries, varying ED amongst 0 (16 ms), 1 (218 ms) and 2 (3267 ms).

Method

As part of a NLP pipeline that aims to identify different aspects of drug usage by patients, one of the atomic steps within this pipeline is the identification of drug names in free text. In this section we describe how string and phonetic similarity metrics can be combined to improve accuracy on identifying misspelt names of drugs within a set of records written in Portuguese. Our approach has two main steps. First, we combine string (*StringSim*) and language-dependent phonetic (*PhoneticMapSim_{PT}*) similarity metrics proposed in [18] in a hybrid similarity

search solution in order to produce a base dictionary of misspelt variations. These metrics were originally designed for the Brazilian Portuguese language. Finally, this dictionary is used as input for a fast Trie-based similarity search algorithm that finds potential candidates to be annotated as drug names in text.

We started using list of 5535 drug names available in the *InfoSaude* system, and searching the most cited drugs in a experimental dataset of clinical notes provided by the *InfoSaude* team (de-identified data with no ethical approval required) from 4748 distinct patients (multiple documents per patient). An exact match search produced a list of 516 drug names, from which the 20 most cited drugs in the text were initially selected (Table 1).

In this first step, we aim to produce a base dictionary of misspelt drug name variations by combining string and phonetic thresholds in order to maximise the accuracy on identifying true positive misspelt words. Such thresholds are used to determine whether a candidate misspelt word correspond to a drug name. Inappropriate low threshold values may return too many false candidates favouring low precision by including words with low similarity values that do not correspond to a drug name. In contrast, high threshold values may exclude possible valid misspelt drug names from the final matching, favouring low recall. The

method used to find the most suitable string and phonetic similarity thresholds is described below:

- We selected a list of candidate words (similar words) for each drug, by finding all words that have at least 3 matching consonantal phonemes in each pair of true positive and candidate drug name or the Edit Distance metric ≤ 3 .
- The returned list of similar words corresponding to a given drug name d was manually analysed. We applied a filter in order to consider candidates words w where $StringSim(d, w) \leq 0.6$ (this threshold can be considered relatively low and resulted approximately 50% of false positive candidates). The final result is a list of 1791 distinct candidate words for the set of drug names listed in Table 1 – an average of 90 similar candidate words per drug.
- The candidate words were manually annotated to identify whether each word corresponds to a valid drug name, resulting 938 positive matches and 853 negative matches. We also automatically annotated each positive and negative match with the corresponding string and phonetic similarity measures (StringSim and PhoneticMapSimPT).
- We used the annotated set of candidates to perform a grid search over the combined string and phonetic similarity values in order to find the best similarity threshold values that favour precision and recall. The list of 20 drugs was split into two groups (10 drugs each) used as training and validation sets. The grid search algorithm is presented in the form of a pseudo-code in Fig. 1.

Table 1 Occurrence (#) of the 20 most cited drug names in a set of 4748 medical records written in Portuguese

Drug name	Number of occurrences
Fluoxetina	18624
Paracetamol	8697
Diazepam	8474
Amitriptilina	8463
Omeprazol	7825
Dipirona	7320
Glicose	5721
Captopril	5383
Insulina	5290
Nimesulida	4228
Clorpromazina	4226
Enalapril	4144
Imipramina	4135
Sinvastatina	3862
Carbamazepina	3853
Amoxicilina	3716
Ibuprofeno	3714
Metformina	3467
Risperidona	3464
Atenolol	3224

The pseudo-code performs an exhaustive search for the best pair of phonetic and string similarity thresholds. The input comprises two manually annotated lists (trainSet and validSet) – containing names of drugs and candidate similar words with the corresponding positive or negative match flag – and a list with 7730 pairs of possible string and phonetic threshold values. 660 pairs of similarity values contain StringSim = 0, i.e. a possible solution considering only the phonetic similarity metric as a threshold. Finally, for each possible pair of threshold values, the algorithm calculates Precision, Recall, and F1 for each set of 10 drugs (trainSet – lines 2-7 – and validSet – lines 8-13). The final thresholds are updated each time both $F1_{train}$ and $F1_{valid}$ simultaneously achieve better values – lines 14-19. After executing the described pseudo-code on the data extracted from the medical record set, we observed a hybrid solution considering both phonetic and similarity thresholds achieved better accuracy on identifying misspelt names of drugs. The hybrid solution combines a smaller phonetic threshold to perform a fast similarity search that result more similar words, coupled with a

```

in:  trainSet, validSet, possibleThresholdSet
out:  phoneticSimilarityThreshold ← 0, stringSimilarityThreshold ← 0
var:  bestF1train ← 0, bestF1valid ← 0;

1:  for (each t in possibleThresholdSet) loop
    // Precision, Recall, F1 for trainSet
2:  TPtrain ← CalcTruePositives(trainSet, t.phoneticSim, t.stringSim);
3:  FPtrain ← CalcFalsePositives(trainSet, t.phoneticSim, t.stringSim);
4:  FNtrain ← CalcFalseNegatives(trainSet, t.phoneticSim, t.stringSim);
5:  Precisiontrain ← TPtrain / (TPtrain + FPtrain);
6:  Recalltrain ← TPtrain / (TPtrain + FNtrain);
7:  F1train ← 2 * (Precisiontrain * Recalltrain) / (Precisiontrain + Recalltrain);
    // Precision, Recall, F1 for validSet
8:  TPvalid ← CalcTruePositives(validSet, t.phoneticSim, t.stringSim);
9:  FPvalid ← CalcFalsePositives(validSet, t.phoneticSim, t.stringSim);
10: FNvalid ← CalcFalseNegatives(validSet, t.phoneticSim, t.stringSim);
11: Precisionvalid ← TPvalid / (TPvalid + FPvalid);
12: Recallvalid ← TPvalid / (TPvalid + FNvalid);
13: F1valid ← 2 * (Precisionvalid * Recallvalid) / (Precisionvalid + Recallvalid);
    // Update Thresholds
14: if (F1train > bestF1train) and (F1valid > bestF1valid) then
15:   bestF1train ← F1train;
16:   bestF1valid ← F1valid;
17:   phoneticSimilarityThreshold ← t.phoneticSim;
18:   stringSimilarityThreshold ← t.stringSim;
19:   endif;
20: end loop;
21: return (phoneticSimilarityThreshold, stringSimilarityThreshold);

```

Fig. 1 A pseudo-code to find similarity thresholds

string similarity threshold that works as a complementary filter. Table 2 depicts the final resulting threshold values.

Results

The final threshold values were used to find positive misspelt names for a broader list of drugs. A total of 1442 misspelt words corresponding to 409 different drug names were identified. Table 3 shows the drug names (except those occurring in the training and validation sets) in

which the greatest number of misspelt forms were found, as well as the corresponding accuracy (precision, recall, F1) on identifying the misspelt variations for each drug. We also compare the accuracy of our approach against the widely used Edit Distance metric.

Information Extraction and NLP systems are traditionally evaluated through precision, recall, and F1-score relevance measures. Precision is equivalent to the amount of retrieved instances that are relevant, while recall is equivalent to the amount of relevant instances that are retrieved. The terms *true positives* (TP) and *true negatives* (TN) represent the correct result and the correct absence of results respectively, while the terms *false positives* (FP) and *false negatives* (FN) correspond to the unexpected result and the missing result respectively. These terms are used to define precision and recall according to Eqs. 1 and 2. In other words, the greater is precision the lesser is the proportion of false positive results, whilst the greater is recall the lesser is the proportion of false negative results. Finally, the F1-score result can be interpreted as the weighted average (or harmonic mean) between precision and recall [24], reaching its best value at 1 and worst score at 0 (Eq. 3).

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

Table 2 Best threshold values found by the grid search method

	Parameter	Value
Training Set	Number of true positives	417
	Number of false positives	31
	Number of false negatives	25
	Precision	0.931
	Recall	0.943
	F1-score	0.937
Validation Set	Number of true positives	477
	Number of false positives	39
	Number of false negatives	19
	Precision	0.924
	Recall	0.961
	F1-score	0.942
Thresholds	Phonetic similarity	0.844
	String similarity	0.831

Table 3 Drugs with the highest number of misspelt variations

Drug name	Number of similar words	Inexact Phonetic Match			F1-Score when using only string match			
		Precision	Recall	F1	$ED \leq 1$	$ED \leq 2$	$ED \leq 3$	$ED \leq 4$
Propranolol	52	0.960	0.979	0.967	0.310	0.819	0.945	0.955
Glibenclamida	49	1.000	1.000	1.000	0.829	0.956	0.955	0.961
Anlodipino	49	0.913	0.976	0.944	0.612	0.938	0.952	0.942
Medroxiprogesterona	47	1.000	0.914	0.955	0.763	0.881	0.955	0.927
Metoclopramida	46	1.000	0.977	0.989	0.750	0.977	0.965	0.964
Loratadina	46	0.837	0.947	0.889	0.774	0.973	0.963	0.955
Dexametasona	45	1.000	0.800	0.889	0.615	0.915	0.954	0.952
Furosemida	43	0.963	1.000	0.981	0.844	1.000	0.976	0.961
Prednisona	42	1.000	0.878	0.935	0.730	0.952	0.976	0.956
Hidroclorotiazida	41	1.000	0.975	0.987	0.776	0.962	0.952	0.940
Diclofenaco	41	0.923	0.947	0.935	0.812	0.914	0.950	0.912
Ciprofloxacino	37	1.000	0.918	0.958	0.520	0.878	0.935	0.922
Espironolactona	36	1.000	1.000	1.000	0.714	0.941	0.948	0.962
Salbutamol	36	1.000	0.972	0.986	0.819	0.956	0.976	0.943
Clonazepam	34	1.000	1.000	1.000	0.692	0.969	0.961	0.939
Beclometasona	33	1.000	0.967	0.984	0.777	0.935	0.961	0.921
Dexclorfeniramina	31	1.000	0.903	0.949	0.708	0.872	0.960	0.959
Metronidazol	30	0.965	0.965	0.965	0.816	0.964	0.942	0.926
Prednisolona	30	0.965	0.965	0.965	0.739	0.925	0.976	0.966
Isossorbida	29	0.963	1.000	0.981	0.761	0.960	0.957	0.936
Average F1-score				0.963	0.718	0.934	0.958	0.945

The best F1 score is highlighted for each drug

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Some drugs reach recall lower than 0.9 (e.g. “*Dexametasona*” and “*Prednisona*”) and “*Loratadina*” has a precision around 0.83, which is lower than most others. Although not being conclusive and still needs further investigation, we found in an initial analysis the observed differences among the scores refer to some prefixes (e.g. “*cap*”, “*clo*”, “*para*”, “*ox*”) and suffixes (e.g. “*mina*”, “*lina*”, “*pina*”, “*tina*”) that are used to compound names of distinct drugs, increasing the value of the similarity scores for negative matches, thus leading to false positives. Some words in Portuguese can also be compound by the verbal derivative form of a noun, such as “*insulinizar*” as a verb referring to the substance “*Insulina*”. All these factors combined increase the probability of a drug name being similar to a more diverse set of distinct words or other drugs in this specific language.

A hybrid solution showed to be efficient on dealing with both phonetic and spelling errors, and combining both string and phonetic similarity thresholds favoured precision and recall when looking for misspelt drug

names. However, this approach suffers in terms of performance in a large corpus. Thus, we used the resulting dictionary of true positive misspelt names of drugs as input for an adapted version of the Trie-based fast search approach algorithm proposed in [22]. This combined approach showed to be efficient (in terms of performance) on finding dictionary-based variations with $\max(ED(word_1; word_2)) = 1$. As a result, hundreds of potential misspelt variations for drug names were identified after processing a new set of medical records comprising approximately 5 million documents. To illustrate the potential use of such combined method, 231 positive misspelt variations for “*Fluoxetina*” (Fluoxetine) and 501 positive misspelt variations for “*Paracetamol*” have been already positively identified. Table 4 shows that some of this variations for the drug “*Fluoxetina*” can have high values for the Edit Distance metric.

Conclusions and future work

In this paper, we presented a hybrid similarity approach that efficiently performs a joint string and language-dependent phonetic similarity search over a set of medical records written in Portuguese. Experimental results showed this method is potentially accurate and able to

Table 4 Examples of misspelt variations for "Fluoxetine" (Fluoxetine) and the corresponding Edit Distance (ED) values

Misspelt variation	ED
dfluoxetine	1
fluoxetine	1
floxetina	1
fluoxetine	1
fluoxetina	1
fluopxetina	1
fluoxertina	1
fluoxetiina	1
fluoxetijna	1
fluoxetin	1
fluoxetinas	1
fluoxetna	1
fluoxetona	1
fluoxettina	1
fluoxetuina	1
fluoxewtina	1
fluoxetina	1
fluozxetina	1
fluuooxetine	1
fluxetina	1
fluyoxetine	1
flhuoxetin	2
flfluoxetin	2
flouxetina	2
fluoxeitna	2
fluoxetian	2
fluxoetina	2
fluxotina	2
fluloextina	3
fluoxetinaate	3
fluoxetinapor	3
flxtina	3
fluoxetinapara	4
inflooxetine	4

identify misspelt names of drugs, overcoming the loss of information inherit from using either exact match search methods or string based similarity search. We coupled the proposed approach with a Trie-based fast similarity search algorithm that is able to use small Edit Distance threshold (≤ 1) over the produced dictionary of misspelt names in order to find a broader number of misspelt variations within an affordable processing time in a large corpus.

Some of the directions in which this work can be extended include: a) adapting the phonetic matching process originally designed to the Portuguese language to be used over large corpora in different languages, such as English; b) integrating our method in a framework for Medical Records Information Extraction applications to address the problem of generically dealing with spelling errors in the information extraction process beyond names of drugs, including other types of clinical variables, such as symptoms and diagnoses; c) exploring the use of machine learning methods to optimally and dynamically tune the threshold parameters and disambiguating misspelt candidates in cases when they are similar to more than one medication; d) comparing the proposed solution with other approximate string match approaches.

Abbreviations

CIS: Outpatient Information System; ED: Edit Distance (or Levenshtein Distance); EHR: Electronic Health Record; HealTAC: Healthcare Text Analytics Conference; ICD: International Code of Diseases; NIH: National Institute for Health Research; RAAL: Ambulatory Care Individual Report

Acknowledgments

We would like to thank the *InfoSaude* team for the permission to use the patient EHR data.

An initial version of this paper has been presented at the Healthcare Text Analytics Conference 2018 (HealTAC), in April 2018 (<http://healtex.org/healtac-2018/>).

About this supplement

This article has been published as part of the *Journal of Biomedical Semantics Volume 10 Supplement 1, 2019: HealTAC-2018: Unlocking Evidence Contained in Healthcare Free-text*. The full contents of the supplement are available online at <https://jbiomedsem.biomedcentral.com/articles/supplements/volume-10-supplement-1>.

Authors' contributions

HT designed the joint string and language-dependent phonetic similarity approach used in the work, performed the experiments and analysed the final results. RD reviewed the analysis and final results. All authors have read and approved the final manuscript.

Funding

This study was funded Health Data Research UK (grant No. LOND1), which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust. ADS is supported by a postdoctoral fellowship from THIS Institute. Publication costs are funded by UCL open access block grant.

Availability of data and materials

The dataset is available at <http://github.com/HeglerTissot/mnd>, including the complete set of drug names and words used in our model, as well as the pre-calculated values for the string and phonetic similarity matching metrics for each pair (*drug, word*).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

- ¹Institute of Health Informatics, University College London, London, UK.
²Health Data Research UK London, University College London, London, UK.
³Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK.

Published: 12 November 2019

References

- Jellouli I, Mohajir ME. An ontology-based approach for web information extraction. In: 2011 Colloquium in Information Science and Technology. IEEE; 2011. <https://doi.org/10.1109/cist.2011.6148583>.
- Pavel S, Euzenat J. Ontology Matching: State of the Art and Future Challenges. *IEEE Trans Knowl Data Eng.* 25(1):158–76. <https://doi.org/10.1109/tkde.2011.253>.
- Karystianis G, Sheppard T, Dixon WG, Nenadic G. Modelling and extraction of variability in free-text medication prescriptions from an anonymised primary care electronic medical record research database. *BMC Med Inf Dec Mak.* 2016;16(1). <https://doi.org/10.1186/s12911-016-0255-x>.
- Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *JAMIA.* 2010;17(5):514–8.
- Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet.* 2012;13(6):395–405. <https://doi.org/10.1038/nrg3208>.
- Senger C, Kaltschmidt J, Schmitt SPW, Pruszydlo MG, Haefeli WE. Misspellings in drug information system queries: Characteristics of drug name spelling errors and strategies for their prevention. *I J Med Inf.* 2010;79(12):832–9.
- Godbole S, Bhattacharya I, Gupta A, Verma A. Building re-usable dictionary repositories for real-world text mining. In: Huang J, Koudas N, Jones GJF, Wu X, Collins-Thompson K, An A, editors. *CIKM.* New York: ACM; 2010. p. 1189–98.
- Levenshtein VI. Binary codes capable of correcting insertions and reversals. *Sov Phys Dokl.* 1966;10(8):707–10.
- Winkler WE. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In: *Proceedings of the Section on Survey Research.* Washington: American Statistical Association; 1990. p. 354–9.
- Stvilia B. A model for ontology quality evaluation. *First Monday.* 2007;12(12). <https://doi.org/10.5210/fm.v12i12.2043>. University of Illinois Libraries.
- Brazilian Ministry of Health: Programa Mais Medicos (More Doctors Program). <http://maismedicos.gov.br/>. Accessed 22 May 2015.
- Bona C. Avaliação de Processos de Software: Um estudo de caso em XP e ICONIX. Master's thesis, Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina (UFSC). 2002.
- Hamming R. Error Detecting and Error Correcting Codes. *Bell Syst Tech J.* 1950;26(2):147–60.
- Tissot H, Peschl G, Fabro MDD. Fast phonetic similarity search over large repositories. In: *Database and Expert Systems Applications - 25th International Conference, DEXA 2014, Munich, Germany, September 1-4, 2014. Proceedings, Part II.* Cham: Springer International Publishing; 2014. p. 74–81.
- Zobel J, Dart P. Phonetic string matching: Lessons from information retrieval. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '96.* New York: ACM; 1996. p. 166–72.
- Droppo J, Acero A. Context dependent phonetic string edit distance for automatic speech recognition. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE; 2010. p. 4358–61. <https://doi.org/10.1109/icassp.2010.5495652>.
- Ladefoged P, Maddieson I. *The Sounds of the World's Languages.* Oxford: Blackwell; 1996.
- Tissot H. Normalisation of imprecise temporal expressions extracted from text. PhD thesis, Federal University of Parana, Brazil, Computer Science Department. 2016.
- Bocek T, Hunt E, Stiller B, Hecht F. Fast similarity search in large dictionaries. Technical Report ifi-2007.02, Department of Informatics, University of Zurich (April 2007). <http://fastss.csg.uzh.ch/>. Accessed 17 Jan 2018.
- Khabsa M, Treeratpituk P, Giles CL. Ackseer: a repository and search engine for automatically extracted acknowledgments from digital libraries. In: *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries.* New York: ACM; 2012. p. 185–94.
- Navarro G. A guided tour to approximate string matching. *ACM Comput Surv.* 2001;33(1):31–88.
- Ji S, Li G, Li C, Feng J. Efficient interactive fuzzy keyword search. In: *Proceedings of the 18th International Conference on World Wide Web, WWW '09.* New York: ACM; 2009. p. 371–80.
- Fenz D, Lange D, Rheinländer A, Naumann F, Leser U. Efficient similarity search in very large string sets. In: Ailamaki A, Bowers S, editors. *Scientific and Statistical Database Management. Lecture Notes in Computer Science, vol 7338.* Berlin: Springer Berlin Heidelberg; 2012. p. 262–79.
- Davis J, Goadrich M. The relationship between precision-recall and roc curves. In: *Proceedings of the 23rd International Conference on Machine Learning, ICML '06.* New York: ACM Press; 2006. p. 233–40. <https://doi.org/10.1145/1143844.1143874>.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

