Evaluating String Distance Metrics for Approximate Dictionary Matching: A Case Study in Serbian Electronic Health Records

Aleksandar Kaplar*, Aleksandra Aleksić*, Milan Stošović**, Radomir Naumović**, Voin Brković**, Aleksandar Kovačević*

* Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia

** Clinic of Nephrology, Clinical Center of Serbia, Belgrade, Serbia

* {aleksandar.kaplar, aleksandra.a, kocha78}@uns.ac.rs

** milanst@eunet.rs, radomirnaumovic450@gmail.com, voin.brkovic@gmail.com

***Abstract*— Matching occurrences of terms (strings) from a dictionary is one of the typical approaches for the extraction of relevant information from large collections of unstructured texts, such as electronic health records (EHRs). A common problem with dictionary-based approaches is matching terms with typographical or orthographical errors. String distance metrics can be used to alleviate this problem, by providing partial matches ranked by the similarity with the sought-out term. In this paper, we evaluate the most commonly used string distance metrics for the task of approximate dictionary matching in EHRs written in the Serbian language. Our results show that the best performing metric is the Jaro distance metric. We also report on the most frequent sources of matching errors.**

I. INTRODUCTION

Electronic health records (EHRs), amongst the structured data, contain a wealth of medical information written in the unstructured form (such as patient disease histories, medications, diagnosis etc.). Extracting information from unstructured parts of EHRs has the potential to improve the care of patients and reveal previously unknown correlations [1].

One of the techniques commonly used for information extraction from large collections of unstructured documents is dictionary matching. The presence of typographical and orthographical errors can significantly reduce the performance of dictionary matching approaches. One solution in dealing with such errors is the use of string distance metrics for approximate string matching [2, 3].

In this paper, our aim is to compare the most commonly used string distance metrics in the context of information extraction from electronic health records written in the Serbian language.

The rest of the paper is organized as follows. Section 2 provides a brief overview of relevant literature. The methodology is presented in Section 3. Section 4 presents the results achieved by the metrics when used on our dataset. Section 5 provides a conclusion for this paper.

II. RELATED WORK

Over the years researchers have studied approximate string matching in the context of different tasks such as probabilistic record linkage [6], identification of common molecular subsequences [7], database record matching [8], and ontology alignment [9].

Cohen et al. conducted a study comparing commonly used string distance metrics for the purpose of name matching tasks [4]. Their results show that tf-idf performed best among token-based distance metrics, whilst among edit-distance metrics Monge-Elkan had the best performance.

Stoilos et al. introduced a new string distance metric and compared it with various edit-distance metrics in the context of ontology alignment [9]. In contrast to [4], their results show that Monge-Elkan and Smith-Waterman were among the worst performing metrics, while Needleman-Wunsch performed well, but wasn't as stable or precise as their proposed metric.

The results of the aforementioned studies illustrate that the choice of the appropriate string distance measures is problem (domain) dependent, and that often researchers need to implement their own custom measures. Additionally, the majority of previous works have only considered dominant and well-researched languages such as English, while the Serbian language remains under-researched in this area. The Serbian language poses a challenge for approximate string matching as it has seven grammatical cases and three grammatical genders (while the English language has neither). In addition, medical texts contain loanwords from languages such as English and Latin that are used in different forms depending on the grammatical case or gender.

Given that the performances of string distance measures are domain dependent, and that medical texts written in the Serbian language are quite challenging for string matching, we decided to conduct our study. To the best of our knowledge, there are no studies for the Serbian language that compare string distance metrics for approximate dictionary matching in the field of medical text mining.

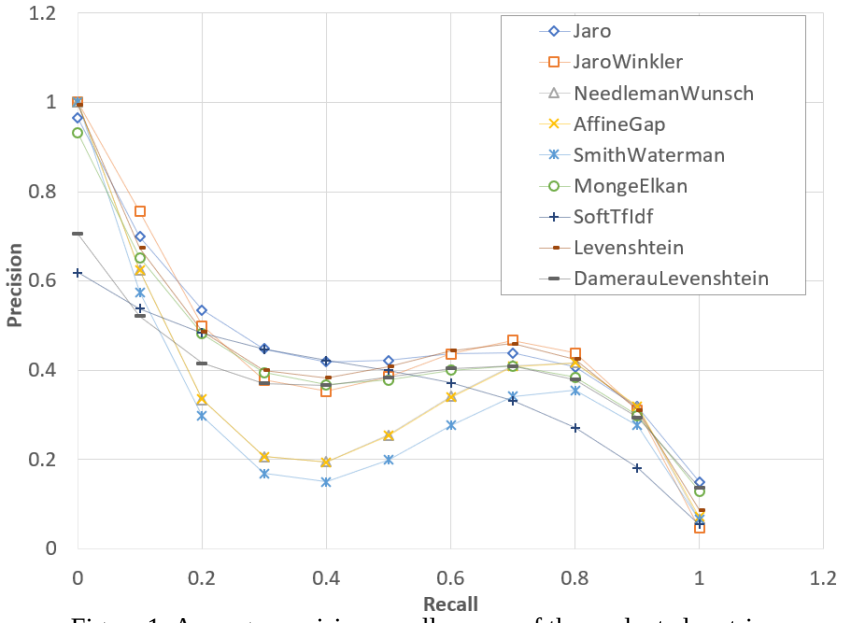

Figure 1. Average precision-recall curves of the evaluated metrics.

III. Methodology

String distances were compared on a corpus of anonymized electronic health records related to assorted nephrological complaints obtained from Clinic of Nephrology at Clinical Center of Serbia.

In the pre-processing stage of our methodology, we have tokenized the corpus and preformed stop-word removal. We note that we did not perform lemmatization or stemming, as the few available stemmers and lemmatizers for the Serbian language, did not perform well on the medical documents that contain a lot of very specific medical terms. As mentioned before, an added difficulty to processing medical corpora written in the Serbian language, besides the fact that it has seven grammatical cases and three grammatical genders, is in the use of loanwords from the English and Latin languages for medical terminology. For example, an English term endoscopy appears in our corpus in the following forms: endoscopksi, endoscopija, endoscopica, endoscopsku, endoscopski, endoscopskim, endoscopom; similarly with the medical term of Latin origin, retroperitoneal: retroperitonealno, retroperitonealnu, retroperitonealni, retroperitonealna, retroperitonealne, retroperitonealis, retroperitonealnom, retroperitonealnog, retroperitoneal-nim, retroperitonealnih.

As the next step in our experiments, we compiled a dictionary containing the top 10 percent of most frequent terms in our corpus. These terms were then matched against our corpus using the following metrics: Jaro [10], Jaro-Winkler [6], Needleman-Wunsch [11], Affine-Gap [12], Smith-Waterman [7], Monge-Elkan [8], SoftTfIdf [4, 5], Levenshtein [13], and Damerau-Levenshtein [14, 13].

Metrics with edit-distance functions, such as Levenshtein, were normalized using the formula (1) to provide a similarity function (a real number in the range [0, 1], where 1 represents two identical terms and 0 two unlike terms) as described in [15].

$$s(x,y) = 1 - \frac{d(x,y)}{\max\left(length(x), length(y)\right)} \quad (1)$$

We used precision, recall, and F1 score as the well-established performance measures to evaluate the metrics. The average precision was calculated as the average value for interpolated precision obtained on 11 evenly spaced recall points as described in [5]. The F1 score is the harmonic mean of recall and precision ($2 \cdot \frac{precision \cdot recall}{precision + recall}$), while the maximal F1 score is the maximum value obtained from precision-recall values of each metric.

IV. Results

Our results show that amongst the worst performing metrics are Needleman-Wunsch, Affine-Gap (which is an extension of Needleman-Wunsch), and Smith-Waterman. Overall the Jaro metric had the best performance, with an average precision of 0.476 and the maximal F1 score of 0.552 (Table 1). The Jaro-Winkler and Levenshtein

Metric	Max F1	Avg.Prec.
Jaro	0.552	0.476
Levenshtein	0.536	0.461
Jaro-Winkler	0.518	0.461
Monge-Elkan	0.518	0.438
Damerau-Levenshtein	0.520	0.398
Needleman-Wunsch	0.478	0.378
Affine-Gap	0.477	0.378
SoftTfIdf	0.453	0.374
Smith-Waterman	0.408	0.336

Table 1. Maximal F1 score and average precision of the compared string distance measures.

Words with severe typographical or orthographical errors		
Matched incorrectly	limfadenopatije	libnfadenopatije
	ekstremiteti	exttremiteti
	fibrinogen	fibronegen
Words with different grammatical case and gender		
Matched incorrectly	alergije	alergijama
	edemu	edemima
	diureza	diureznog

Table 2. Representative terms for matching error categories.

metrics also achieved high maximal F1 scores of 0.518 and 0.536 respectively.

The average values for interpolated precision on 11 evenly spaced recall points are presented in Figure 1. As show in the figure, we can see that even though the Jaro metric had the best performance on our dataset, it didn't significantly outperform other well performing metrics.

During the evaluation of the results we've determined that the tested metrics produced a high number of false positive and false negative matches. And as such, they wouldn't be fully suited for dictionary matching task on our dataset.

During the analysis of the results, we determined that most of the matching errors can be grouped into two categories. The first category are the words with severe typographical or orthographical errors, and the second category are the words with different grammatical case and gender. Table 2 presents representative terms for both matching error categories. We suspect that most of the matching errors of the second category could be solved with the use of an appropriate lemmatizer, while the errors in the first category would require a new string matching metric.

V. Conclusion

In this paper we have analyzed the performance of most commonly used string matching metrics in the context of information extraction from electronic health records written in the Serbian language.

Our results show that amongst the worst performing metrics were Needleman-Wunsch, Affine-Gap, and Smith-Waterman, On the other hand, the results also show that the metric with the best performance, the Jaro metric, didn't significantly outperform other well performing metrics.

From our result analysis, we can conclude that in order to achieve the best performance for dictionary matching task on our dataset a new string distance metric and lemmatizer for the Serbian language, with included medical terminology, needs to be developed.

Acknowledgment

This work has been partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (projects III-44010 and III-47003).

References

[1] Jensen, Peter B., Lars J. Jensen, and Søren Brunak. "Mining electronic health records: towards better research applications and clinical care." Nature Reviews Genetics 13.6 (2012): 395.

[2] Wang, Wei, et al. "Efficient approximate entity extraction with edit distance constraints." Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. ACM, 2009.

[3] Sauleau, Erik A., Jean-Philippe Paumier, and Antoine Buemi. "Medical record linkage in health information systems by approximate string matching and clustering." BMC medical informatics and decision making 5.1 (2005): 32.

[4] Cohen, William, Pradeep Ravikumar, and Stephen Fienberg. "A comparison of string metrics for matching names and records." Kdd workshop on data cleaning and object consolidation. Vol. 3. 2003.

[5] Bilenko, Mikhail, et al. "Adaptive name matching in information integration." IEEE Intelligent Systems 18.5 (2003): 16-23.

[6] Winkler, William E. "The state of record linkage and current research problems." *Statistical Research Division, US Census Bureau*. 1999.

[7] Waterman, M. S. "Identification of common molecular subsequence." *Mol. Biol* 147 (1981): 195-197.

[8] Monge, Alvaro E., and Charles Elkan. "The Field Matching Problem: Algorithms and Applications." *KDD*. 1996.

[9] Stoilos, Giorgos, Giorgos Stamou, and Stefanos Kollias. "A string metric for ontology alignment." *International Semantic Web Conference*. Springer, Berlin, Heidelberg, 2005.

[10] Jaro, Matthew A. "Probabilistic linkage of large public health data files." Statistics in medicine 14.5-7 (1995): 491-498.

[11] Needleman, Saul B., and Christian D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins." Journal of molecular biology 48.3 (1970): 443-453.

[12] Waterman, Michael S., Temple F. Smith, and William A. Beyer. "Some biological sequence metrics." Advances in Mathematics 20.3 (1976): 367-387.

[13] Levenshtein, Vladimir I. "Binary codes capable of correcting deletions, insertions, and reversals." Soviet physics doklady. Vol. 10. No. 8. 1966.

[14] Damerau, Fred J. "A technique for computer detection and correction of spelling errors." Communications of the ACM 7.3 (1964): 171-176.

[15] Doan, AnHai, Alon Halevy, and Zachary Ives. Principles of data integration. Elsevier, 2012.