Named Entity Discovery for the Drug Domain

Nasi Jofche
Faculty of Comp. Sci. and Eng.
Ss. Cyril and Methodius University
Skopje, North Macedonia
nasi.jofche@finki.ukim.mk

Milos Jovanovik
Faculty of Comp. Sci. and Eng.
Ss. Cyril and Methodius University
Skopje, North Macedonia
milos.jovanovik@finki.ukim.mk

Dimitar Trajanov
Faculty of Comp. Sci. and Eng.
Ss. Cyril and Methodius University
Skopje, North Macedonia
dimitar.trajanov@finki.ukim.mk

***Abstract*—Medical datasets that contain data relating to drugs and chemical substances, in general tend to contain multiple variations of a generic name which denotes the same drug or a drug product. This ambiguity lies in the fact that a single drug, referenced by a unique code, has an active substance which can be known under different chemical names in different countries, thus forming an obstacle during the process for extracting relevant and useful information. To overcome the issues presented by this ambiguity, we developed a scalable, term frequency based data cleaning algorithm, that solely uses the data available in the dataset to infer the correct generic name for each drug based on text similarities, thus forming the roots for building a model that would be able to predict generic names for related and previously unseen drug records with high accuracy. This paper describes the application of the algorithm towards the cleaning and standardization process of an already populated drug products availability dataset, by representing all of the variations of a substance under a single generic name, thus eliminating ambiguity. Our proposed algorithm is also evaluated against a Linked Data approach for detecting related drug products in the dataset.**

***Index Terms*—Named Entity, Data Cleaning, Text Similarity, Drug Data, Drugs.**

I. Introduction

The drug product availability dataset, known as Linked-Drugs and available through the Global Open Drug Dataset (GODD) application [3, 21], contains crawled data on drug products registered in a large set of countries around the world, along with information about their respective active substances [4] and unique ATC codes [2]. The dataset, due to its nature, is prone to ambiguity related to multiple variations for a single active substance for a drug product. This ambiguity refers to the fact that a single drug might have an active substance – also referred to as *generic name* – which is known under a different name in different countries – e.g. the chemical substance Paracetamol is also known as Paracetamolum, Acetaminophen, N-acetyl-para-aminophenol, etc. [5], all categorized under the same unique ATC code. This ambiguity in the dataset might lead to disorienting results while trying to extract useful information from the data, thus can present an obstacle that needs to be eliminated during the data preprocessing phase [14]. This is important because different drug products can be labeled as *related* if they have the same active substance (generic name). Then, these related drug products can be used as alternatives of each other in various real-world cases. Therefore, the ability to correctly identify related drug products even when their active substances are not present in the dataset, or have non-matching values, is of high importance.

Besides the application of the widely known steps related to the data preprocessing phase for a dataset, such as handling missing values [11] which is a common scenario in our case, specific drug domain-based preprocessing and cleaning rules are required as well.

This paper presents a preprocessing algorithm that solely uses the available data from the dataset to infer the correct active substance (generic name) for a drug, thus reducing ambiguity. Since the dataset has a significant amount of missing values for the active substances of the drug products, the algorithm is based on a combination of text similarity metrics [15, 17] and ATC code equality. The algorithm consolidates the active substances of all groups of related drug products, by setting them to or replacing them with the most common text variant of the active substance of the given group of drug products.

For the purpose of testing the algorithm, two text similarity techniques were used: cosine similarity and Levenshtein distance. The algorithm is scalable and supports usage of other text distance metrics, as well. Its purpose is to provide preliminary results that build the roots of a future model that can be used to further discover named entities from previously unseen drug related data. The thorough algorithm description, as well as the obtained results are given in the following sections. Finally, the accuracy of our algorithm is tested against an entity detection approach based on Linked Data that detects all similar drug names in order to extract the active substance (generic name).

II. Related Work

Multiple efforts have been made into assessing the drug availability across different countries under consolidated datasets. The analysis in [9] shows a specific subset of drugs and their availability throughout different countries in Europe. On the other hand, the analysis made by [26] indicates the need for increased availability of drugs in 11 countries of the Asia Pacific Region.

The efforts made by [19, 20] use the Linked Data approach to consolidate drug product data in Macedonia, and then on a global scale [21]. These approaches provide a global overview of the drug products which are registered and sold in different countries, and the provide the ability to identify and analyze

related drug products across and between countries. We use some of the approaches described in these papers to compare the accuracy of our proposed algorithm, later in this paper.

These efforts, as well as other similar research works, such as [16], present solutions based on Linked Data that reduce the ambiguity related to drugs branded in multiple names, overcoming the named entity ambiguity obstacle.

Other previous research efforts are focused on named entity recognition and detection from plain text or different datasets. Such different approaches for named entity recognition are described in [13, 23, 24], leading to diverse entity detection approaches in regard to the business domain [10, 18, 22, 25]. The different techniques analyzed for named entity detection are tightly coupled to the domain that they are applied to.

On the other hand, recent advances in this field indicate high accuracy while applying entity detection and linking algorithms on language independent datasets, as discussed in [12]. This proposed algorithm can be trained on one language and is shown to perform well on other languages without any change. It achieves the entity linking based on different kinds of language independent features and a discriminative ranking function.

III. Generic Names Dataset

In order to assess the accuracy of our algorithm, we created a dataset of generic names by extracting drug related information from DBpedia [1]. This dataset contains information about generic names for drugs, i.e. active substances, accompanied by the respective ATC codes in the following format:

Generic Name	ATC Code	Similar Substances
Paracetamol	N02BE01	Acephen:Acetaminophen...

The information in the dataset was gathered by querying the DBpedia endpoint using a custom SPARQL [6] query, focusing on the wikiPageRedirects property [7] to extract all different names under which a substance is known. The query returned 1,675 records which we used to test the accuracy of our algorithm, i.e. its ability to detect the correct generic names.

IV. Algorithm Application & Results

The first step in the process of the dataset analysis was focused on drug product groupings by their respective countries, in order to extract the information for the total number of initial generic names, both correct and incorrect. The dataset of generic names was used to assess this accuracy, by comparing the active substance (generic name) value of each drug product from our LinkedDrugs dataset to the full list of generic names. Table I gives an overview on the initially correct generic names for the respective countries, by analyzing a total of 10 countries.

Due to the large size of the version of the LinkedDrugs dataset being analyzed – 125.424 drug product records at the time of analysis – and the complexity of the algorithm being O(n^2), we decided to assess the algorithm accuracy

TABLE I
Initial Generic Name Accuracy of Drug Products by Country

Country	Incorrect	Correct	Correct [%]
DK	6,410	2,535	28.33
BIH	2,935	1,506	33.91
AZ	2,913	1,529	34.42
US	13,255	7,159	35.06
SL	1,468	933	38.85
BE	4,554	4,263	48.34
MT	3,193	3,506	52.33
CY	1,951	2,587	57.00
FIN	2,641	3,595	57.64
MK	983	2,378	70.75

based on smaller subsets chosen randomly. The algorithm was applied using both cosine similarity and Levenshtein distance, in combination with the available ATC codes for identifying similar substances. Each drug product of the dataset was compared respectively by measuring text distance and only counting distances above a specified threshold, thus forming a sparse vector with a length same as the selected subset. In addition to that, each drug product with the same ATC code was also analyzed to create an additional sparse vector, which was combined in a union relationship with the frequency vector. The resulting vector was used to replace each entity with the most frequent one, thus detecting the most common generic name for the substance which is the active ingredient for the given group of drug products. The resulting cleaned dataset records were compared to the dataset of generic names, to assess the generic name recognition accuracy. The obtained results are given in Fig. 1 and Fig. 2, using cosine and Levenshtein metrics respectively, which indicate an increase in accuracy as the subset size increases. For each subset size we chose five random subsets, and then averaged the results.

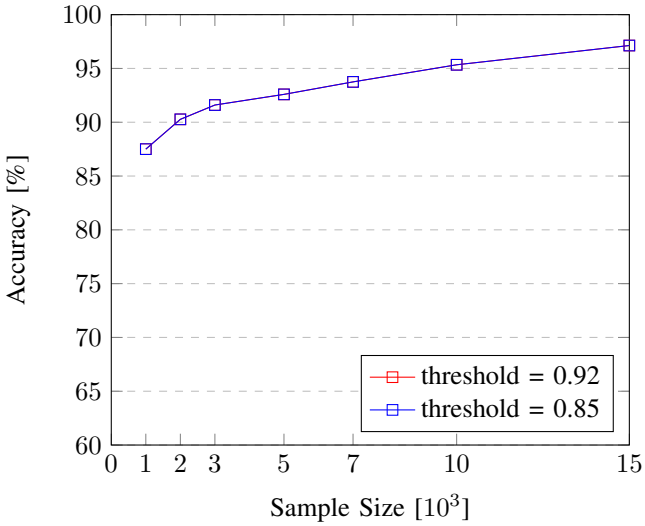

Fig 1. Accuracy Assessment using Cosine Similarity

The obtained results indicate high accuracy, which is attributed to the selected subsets containing mixed data from different countries. We can conclude that both similarity

thresholds that were chosen while assessing the cosine similarity performance, show the same results.

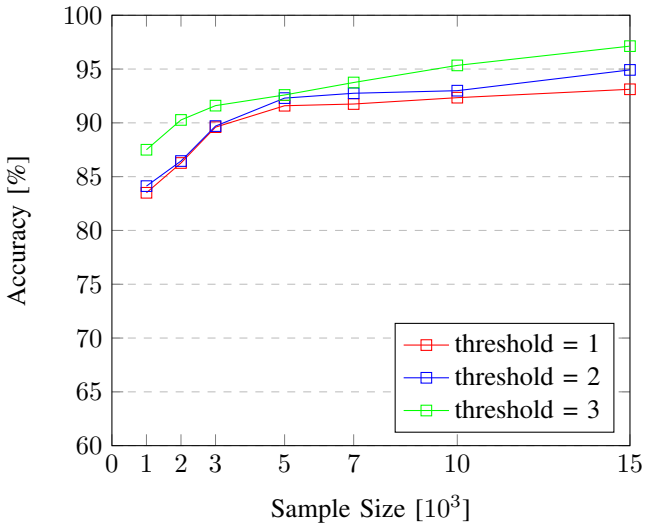

Fig 2. Accuracy Assessment using Levenshtein Distance

The algorithm's performance using the Levenshtein distance metrics with variable distance thresholds indicates high accuracy as well. It is worth noting that for smaller data subsets and a smaller threshold, the obtained accuracy was slightly lower – a result related to the missing ATC code values for drug products. While using the distance threshold = 3, the obtained accuracy was the same as the accuracy obtained while using the cosine similarity.

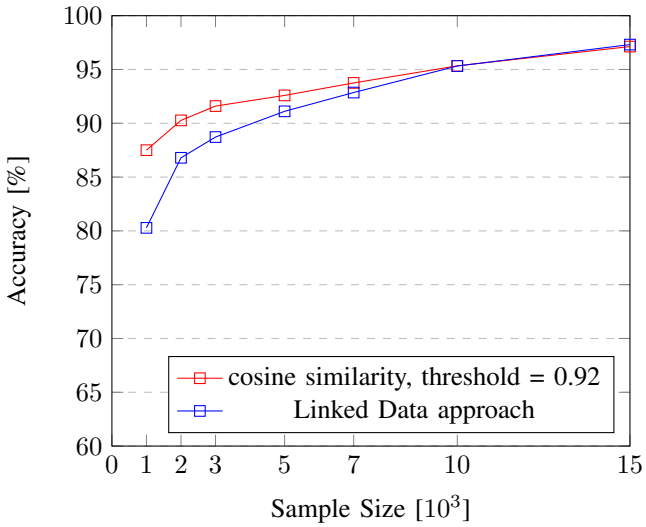

Fig 3. Comparison to the Linked Data Approach

V. Comparison to the Linked Data Approach

In this section we evaluate the accuracy of our proposed algorithm against the Linked Data approach for finding similar drug names in the dataset [21]. We use OpenRefine [8] to apply a transformation script to our dataset, in order to identify similar drugs and infer the generic name by reconciling against drug entities from DBpedia. The obtained results are compared to the results obtained from our algorithm by using the cosine similarity with similarity threshold = 0.92, and are given in the chart in Fig. 3. These results indicate that our algorithm outperforms the Linked Data approach for smaller dataset subsets, while it is slightly outperformed for larger subsets.

VI. Future Work

The described algorithm is part of an ongoing research, serving as the basis for a highly accurate model capable of predicting the generic names of drug products from data that it has not encountered before, with a sole purpose of extracting useful information from the dataset and using the obtained information for making country-related predictions regarding the drug availability.

After this highly accurate, domain-specific cleaning process, the next step is building and improving a neural network that will correctly classify the incoming data from the crawlers. Since the crawlers used in the GODD applications are continuously enriching the dataset, applying the algorithm to the entire dataset every time a new record gets extracted is inefficient, thus a neural network that will be trained using the cleaned and accurate data obtained from the algorithm would be a highly scalable solution.

The same algorithm can further be used to detect company name entities, as well, which are also available in the dataset under different names.

VII. Conclusion

Besides the challenges faced when applying the common cleaning steps to a dataset, applying domain-specific cleaning techniques is a challenge of its own. In this paper we presented a highly accurate algorithm focused on cleaning a drug product availability dataset by detecting named entities and standardizing generic names of substances across all records. The described algorithm uses text similarity metrics: cosine similarity and Levenshtein distance, in combination with the available ATC codes from the dataset. The results were assessed using variable similarity and distance thresholds, leading to high accuracy as the subset size increases. In the end, our algorithm was tested against a Linked Data approach that detects similar drugs while querying the DBpedia knowledge graph. The Linked Data approach was shown to perform slightly better as the data subset size increased, while it was outperformed by our proposed algorithm for small sized datasets.

Acknowledgment

The work in this paper was partially financed by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje.

References

[1] Dbpedia. Accessed 08 Apr 2019. http://dbpedia.org/.
[2] ATC Codes: Structure and Principles. Accessed 09 Apr 2019. http://www.whocc.no/atc/structure_and_principles.
[3] Global Open Drug Dataset. Accessed 12 Apr 2019. http://godd.finki.ukim.mk/.

[4] Active Substance. Accessed 13 Apr 2019. https://www.ema.europa.eu/en/glossary/active-substance.

[5] Paracetamol Brand Names. Accessed 13 Apr 2019. https://adf.org.au/drug-facts/paracetamol/.

[6] SPARQL. Accessed 13 Apr 2019. https://www.w3.org/TR/rdf-sparql-query/.

[7] Wiki Page Redirects. Accessed 13 Apr 2019. http://dbpedia.org/ontology/wikiPageRedirects.

[8] OpenRefine. Accessed 24 Apr 2019. http://openrefine.org/.

[9] A. Baftiu, C. Johannessen Landmark, V. Nikaj, I.-L. Neslein, S. I. Johannessen, and E. Perucca. Availability of Antiepileptic Drugs Across Europe. *Epilepsia*, 56(12):e191–e197, 2015.

[10] C. Brun and C. Hagege. Semantic Compatibility Checking for Automatic Correction and Discovery of Named Entities, Aug. 16 2011. US Patent 8,000,956.

[11] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang. Data Cleaning: Overview and Emerging Challenges. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2201–2206. ACM, 2016.

[12] L. Ding and B. Dong. Language Independent Entity Linking. pages 724–729, 11 2018.

[13] S. Eltyeb and N. Salim. Chemical Named Entities Recognition: A Review on Approaches and Applications. *Journal of cheminformatics*, 6(1):17, 2014.

[14] S. García, J. Luengo, and F. Herrera. *Data Preprocessing in Data Mining*. Springer, 2015.

[15] W. H. Gomaa and A. A. Fahmy. A Survey of Text Similarity Approaches. *International Journal of Computer Applications*, 68(13):13–18, 2013.

[16] A. Hasnain, M. R. Kamdar, P. Hasapis, D. Zeginis, C. N. Warren, H. F. Deus, D. Ntalaperas, K. Tarabanis, M. Mehdi, and S. Decker. Linked Biomedical Dataspace: Lessons Learned Integrating Data for Drug Discovery. In *International Semantic Web Conference*, pages 114–130. Springer, 2014.

[17] A. Huang. Similarity Measures for Text Document Clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56, 2008.

[18] U. Irmak and R. Kraft. Scalable Semi-Structured Named Entity Detection, 2011. US Patent 8,073,877.

[19] M. Jovanovik, B. Najdenov, G. Strezoski, and D. Trajanov. Linked Open Data for Medical Institutions and Drug Availability Lists in Macedonia. In *New Trends in Database and Information Systems II*, pages 245–256. Springer, 2015.

[20] M. Jovanovik, B. Najdenov, and D. Trajanov. Linked Open Drug Data from the Health Insurance Fund of Macedonia. In *10th Conference for Informatics and Information Technology (CIIT)*, 2013.

[21] M. Jovanovik and D. Trajanov. Consolidating Drug Data on a Global Scale Using Linked Data. *Journal of Biomedical Semantics*, 8(1):3, 2017.

[22] K. Li, Y. Li, Y. Zhou, Z. Lv, and Y. Cao. Knowledge-Based Entity Detection and Disambiguation, 2017. US Patent 9,665,643.

[23] S. Liu, B. Tang, Q. Chen, and X. Wang. Drug Name Recognition: Approaches and Resources. *Information*, 6(4):790–810, 2015.

[24] D. Nadeau and S. Sekine. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[25] J. D. Rennie and T. Jaakkola. Using Term Informativeness for Named Entity Detection. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 353–360. ACM, 2005.

[26] H. Wang, Q. Sun, A. Vitry, and T. A. Nguyen. Availability, Price, and Affordability of Selected Essential Medicines for Chronic Diseases in 11 Countries of the Asia Pacific Region: A Secondary Analysis. *Asia Pacific Journal of Public Health*, 29(4):268–277, 2017.