German Medical Data Sciences: Bringing Data to Life
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210073

First Steps to Evaluate an NLP Tool's Medication Extraction Accuracy from Discharge Letters

Deniz CALISKAN[a1], Jakob ZIERK[a,b], Detlef KRASKA[a], Stefan SCHULZ[c], Philipp DAUMKE[c], Hans-Ulrich PROKOSCH[a,d], and Lorenz A. KAPSNER[a]

[a]*Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Erlangen, Germany*
[b]*Department of Pediatrics and Adolescent Medicine, Universitätsklinikum Erlangen, Erlangen, Germany*
[c]*Averbis GmbH, Freiburg, Germany*
[d]*Chair of Medical Informatics, Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Erlangen, Germany*

Abstract. Introduction: The aim of this study is to evaluate the use of a natural language processing (NLP) software to extract medication statements from unstructured medical discharge letters. Methods: Ten randomly selected discharge letters were extracted from the data warehouse of the University Hospital Erlangen (UHE) and manually annotated to create a gold standard. The AHD NLP tool, provided by MIRACUM's industry partner was used to annotate these discharge letters. Annotations by the NLP tool where then compared to the gold standard on two levels: phrase precision (whether or not the whole medication statement has been identified correctly) and token precision (whether or not the medication name has been identified correctly within correctly discovered medication phrases). Results: The NLP tool detected medication related phrases with an overall F-measure of 0.852. The medication name has been identified correctly with an overall F-measure of 0.936. Discussion: This proof-of-concept study is a first step towards an automated scalable evaluation system for MIRACUM's industry partner's NLP tool by using a gold standard. Medication phrases and names have been correctly identified in most cases by the NLP system. Future effort needs to be put into extending and validating the gold standard.

Keywords. Natural Language Processing, Data Analysis, Patient Discharge Summary, Electronic Data Processing

1. Introduction

Hospital admission and discharge are transition points in healthcare that rely on written communication between physicians. The written communication in form of discharge letters often consist of unstructured narrative text. These unstructured clinical records contain a high amount of medical data that is complimentary to structured records [1].

[1] Corresponding Author: Deniz CALISKAN, deniz.caliskan@uk-erlangen.de; Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Krankenhausstraße 12, 91054 Erlangen, Germany

Due to the high amount of narrative information presented in a discharge letter, it can be challenging to extract relevant data for medical research and data analysis from unstructured parts of clinical records [2].

Based on this need, the *Third i2b2 Workshop on Natural Language Processing Challenges for Clinical Records* in 2010 concentrated on identifying medications and corresponding relevant information, such as dosages, frequencies and treatment duration, in clinical records [3]. This workshop demonstrated that it is indeed possible to extract structured data from clinical narratives, even though the state-of-the-art software at that time was not capable of easily extracting all necessary information. Almost a decade later, Natural Language Processing (NLP) has improved significantly and commercial systems that produce structure extracts from unstructured parts of clinical records are available.

Since 2018, the German Medical Informatics Initiative (MII) [4] is funded by the German Federal Ministry of Education and Research (BMBF) with the goal to overcome the challenges of digitalization in medicine. The MIRACUM consortium [5], one of four consortia funded within MII, aims at tapping hospital routine data and integrating them into research data repositories, such as *i2b2* [6] and *OMOP* [7,8], in order to make them available for scientific purposes.

Clinical records are of special interest, as they contain a high amount of data [9]. Medical records commonly include considerable amounts of unstructured text documents, which make data extraction and re-use challenging [2]. In order to address this aspect, the MIRACUM consortium has teamed up with an industry partner that offers an NLP system, specialized in the medical field. For high quality research it is very important to assess completeness and accuracy of data extracts obtained by the NLP system.

In this study, we present the steps towards setting up an automated pipeline in order to evaluate the accuracy of this NLP system in terms of extracting medication information from unstructured medical discharge letters.

2. Study context

A large amount of information in medical records is based on unstructured narrative text [9]. When using NLP software to extract details from e.g. discharge letters at a large scale, the full manual verification of the correctness of the extracted information is nearly impossible. Within MIRACUM, we plan to provide future research with background information on the quality of the data extracted from medical discharge letters. Therefore we aim at estimating the correctness of the NLP tool.

This proof-of-concept study is conducted at the University Hospital Erlangen (UHE), a German University Hospital with appx. 1,400 beds and 65,400 in-patient treatments in 2018 [10]. Over half a million discharge letters since 2007 are already available via UHE's data warehouse (DWH). Even more letters could be made accessible by digitalizing letters from the archives.

3. Methods

3.1. Study design

The study is a retrospective observational accuracy study without intervention and was approved by the local ethics committee (Ethik-Kommission der Friedrich-Alexander-Universität Erlangen-Nürnberg, No. 207_20 Bc).

3.2. Theoretical background of the study

Uzuner et al. describe seven so-called information 'fields' regarding prescribed medication that can be extracted from discharge summaries: medication, dosage, mode, frequency, duration, reason and list/narrative [3]. The latter indicates if the information was extracted from some structured part (list) or from some unstructured text within the discharge letter.

While all of these fields are required to extensively describe the medication of a patient, we focused here on the field 'medication' only, in order to set up an exemplary automated analysis and evaluation pipeline, which can be scaled up and adapted to other medication related entities as well as further medical concepts in the future.

3.3. Study pipeline

This study aims to evaluate the correctness of an NLP system in terms of detecting the 'medication' field in ten clinical discharge summaries. To achieve that, three steps were necessary. At first, ten randomly selected discharge letters, dated between 2007 and 2019, were extracted from the UHE DWH using a Python extract-transform-load (ETL) script. A gold standard was created by manually annotating these letters with the following approach: In every discharge letter, all medication-related phrases have been identified by their corresponding starting and ending positions inside the text document. A 'medication phrase' includes at least the 'medication' field plus any of the fields 'medication', 'dosage', 'mode', 'frequency', 'duration' and 'reason' that semantically belong to the corresponding 'medication' field. The 'reason' field can also include negating words such as 'stopped' or 'is allergic to'. For example the phrase in the sentence "The patient has no known allergies to Ibuprofen" would be "no known allergies to Ibuprofen" and the phrase in "The patient takes Ibuprofen 100mg 1-1-1 daily for two weeks" would be "Ibuprofen 100mg 1-1-1 daily for two weeks". The position is measured from the first character of the first field to the last character of the last corresponding field and presented by their absolute character positions within the text document. One distinct medication can occur multiple times in the same discharge letter but with different starting and ending positions. A 'medication token' is the corresponding medication field in a medication phrase. The gold standard consists of the starting and ending positions of the identified phrases and the corresponding medication field.

Next, the raw text-based discharge letters were submitted to the NLP tool. The tool provides by default a 'discharge-pipeline', which annotates starting and ending positions of the complete medication phrase as well as all discovered 'fields'.

Furthermore, laboratory sections were also identified and marked within the gold standard. Medication names mentioned within these laboratory sections, e.g. the reporting of drug levels, were excluded from the NLP tool's results prior to the

subsequent analyses, if they have been correctly recognized as belonging to a laboratory section by the NLP system.

In the final step, the results obtained using the NLP system were compared to the gold standard.

3.4. Outcome measures

The outcome measures reported by Uzuner et al. [3], *viz.* 'precision', 'recall' and 'F-measure', have been adapted to our evaluation pipeline. Similar to their approach, these three measures are evaluated on a phrase level and a token level, for each discharge letter. Finally, for each measure the mean across all discharge letters is reported.

In order to calculate these measures, the results of the comparison between the annotation by the NLP system and the gold standard are coded by binary values: "0" = NLP tool results does not match gold standard; "1" = NLP tool result matches gold standard. Furthermore, measures at token level are calculated only if the phrase has been identified correctly.

The evaluation workflow is depicted in Figure 1. At first, each starting position of the NLP tool's result is compared to each starting position of the gold standard. A phrase is correct and coded as "1", if the starting positions exactly match. Only if this condition is true, it is examined in a second step whether the NLP tool discovered the corresponding medication token correctly. This token is considered as correct and coded as "1", if the medication name identified by the NLP tool exactly matches the definition in the gold standard.

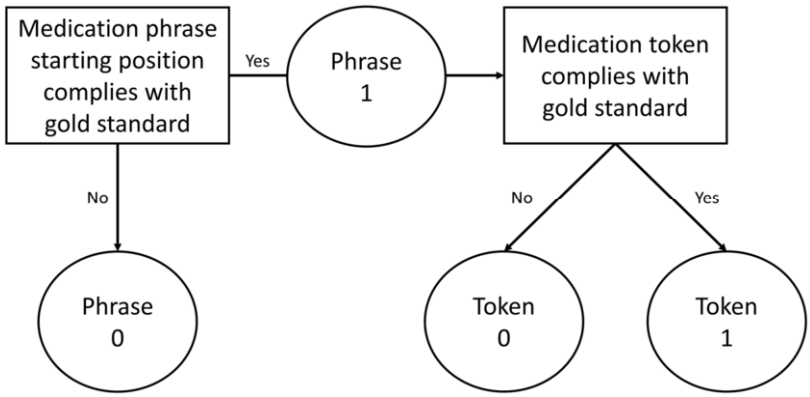

Figure 1. Evaluation Workflow: At first, the starting position of every phrase recognized by the NLP system is compared with the gold standard. If these positions match, it is further checked, if the medication has been correctly identified within the recognized phrase. The results are coded by binary values: "1" = NLP tool result matches gold standard; "0" = NLP tool results does not match gold standard.

3.5. Methods for data analysis

The NLP tool *Averbis Health Discovery* (AHD), Version 5.22.0 [11], provided by the MIRACUM consortium's industry partner *Averbis GmbH*, has been used to automatically annotate the discharge letters.

The discharge letters were extracted from the DWH using the structured query language (SQL) and Python 3.6 [12].

All letters were pre-processed by extracting the text body from the XML structure and transformed into a human readable text file. Those pre-processed raw texts were subsequently used to manually annotate the gold standard on the one hand and to provide them to the REST API of the NLP software by using Python scripts, on the other hand. The analysis of the results and the computation of the evaluation metrics was done by using the Python programming language and R [13].

4. Results

The average length of discharge letters under investigation was 6,838 characters with an average of 13 medication phrases in the gold standard. Table 1 presents the number of annotations covered by the gold standard with the annotations determined using the NLP system, the corresponding document lengths and the resulting evaluation metrics.

The overall average phrase level precision was 0.935, the average phrase recall was 0.806, and the phrase level F-measure was 0.852. The phrase level F-measure varied between 0.5 and 1 with a standard deviation of 0.157.

Token level precision, recall and F-measure were only computed, when the corresponding phrase has been identified correctly. Therefore, the total number of tokens taken into consideration is equal or smaller than the total number of phrases. On the token level, the overall average precision was 0.969, the average recall was 0.909, and the average F-measure was 0.936. The phrase level F-measure varied between 0.783 and 1 with a standard deviation of 0.085.

Table 1. Evaluation metrics: Phrase and token level precision, recall and F-measure values for each discharge letter. P: phrase; T: token; PR: precision; RE: recall; F1: F-measure; DOC-L: document length; N: number of annotations; GS: gold standard; NLP: NLP (natural language processing) system.

Letter	N (GS/NLP)	DOC-L	PR (P)	RE (P)	F1 (P)	PR (T)	RE (T)	F1 (T)
1	18 / 16	10514	1.000	0.889	0.941	0.938	0.938	0.938
2	21 / 22	7111	0.909	0.952	0.930	1.000	0.909	0.952
3	12 / 11	5042	1.000	0.917	0.957	1.000	1.000	1.000
4	15 / 5	7961	1.000	0.333	0.500	1.000	1.000	1.000
5	7 / 7	4149	1.000	1.000	1.000	1.000	1.000	1.000
6	15 / 13	14595	0.769	0.667	0.714	0.900	0.692	0.783
7	14 / 13	4620	1.000	0.929	0.963	1.000	1.000	1.000
8	12 / 10	4938	1.000	0.833	0.909	1.000	1.000	1.000
9	6 / 5	5037	0.800	0.667	0.727	1.000	0.800	0.889
10	8 / 8	4411	0.875	0.875	0.875	0.857	0.750	0.800
average	12.8 / 11	6838	0.935	0.806	0.852	0.969	0.909	0.936

5. Lessons learned

The NLP tool detected medication related phrases with an overall F-measure of 0.852. The medication name was correctly identified with an overall F-measure of 0.936.

The results show that the information extracted by the AHD NLP system from semi-structured medical discharge letters can be evaluated automatically by using a gold standard. The NLP system achieved high token precision and recall values, suggesting that it is able to identify medication names correctly.

Detection rates depend heavily on the gold standard definition. I.e., multiple (correct) definitions of what is a medication phrases can exist [3]. Even small discrepancies between the phrase definition in the gold standard and the NLP tool's result can have large impacts on the outcome measures.

For example, if the phrase is "5mg Ibuprofen per day" and the gold standard definition only marks the medication or ingredient name, the phrase would start at "Ibuprofen" in the gold standard but the tool would correctly identify the dose statement and start at "5mg". That would then lead to not matching starting positions and thus, a poorer phrase precision and recall.

When interpreting our results, it should be taken into consideration that during the design of our study, the creation and definition of the gold standard has been aligned to the NLP tool's output for technical reasons, i.e. to be able to meaningfully compare them. Furthermore, the gold standard annotation has not been validated by multiple annotators.

The large extension of the gold standard by the annotation of more discharge letters from various clinical departments at UHE, implemented by multiple experts, will be the focus of our future efforts. We also aim at distinguishing whether the medication statement related information was extracted from a narrative part or a list within the discharge letter. This will help to better understand the capabilities of the NLP system to detect medical concepts from several disciplines and will likely reduce the impact of uncertainties, naturally introduced in narratives. Furthermore, we will extend the herewith gathered experience to all other fields that occur in medication statements as well as transfer this knowledge also to further topics covered by discharge letters.

Furthermore, it should also be mentioned that in this proof-of-concept study only the correct matching of the starting positions has been used as a criterion for successfully detecting a medication phrase. A future improvement could include e.g. the introduction of fuzziness to a certain degree, when comparing also ending positions.

6. Conclusion

This work is a first step for moving towards an automated scalable evaluation system of NLP software by using a gold standard. Although the results of this proof-of-concept study are not generalizable due to analyzing only a small amount of discharge letters from only one clinical department, we were able to establish an automated evaluation pipeline at UHE that can be used to accompany future ETL processes when extracting narratives at larger scale. This is necessary for researchers to estimate the 'truthfulness' of the extracted information.

Declarations

Authors' contribution: Conceptualization: LAK, DC; Data curation: DC; Formal analysis: DC; Investigation: DC, LAK; Methodology: LAK, DC; Project administration: LAK, HUP, DK; Resources: HUP; Supervision: LAK, HUP, DK; Validation: LAK, JZ;

Visualization: DC, LAK; Writing - original draft: DC, LAK; Writing - review and editing: DC, LAK, HUP, JZ, SS, PD, DK.

Acknowledgements: This work was funded in part by the German Federal Ministry of Education and Research (BMBF) within the Medical Informatics Initiative (MIRACUM Consortium) under the Funding Number FKZ: 01ZZ1801A. The present work was performed in (partial) fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum." of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (DC).

Conflict of Interest: All authors state that they have no conflict of interests. Averbis GmbH has nothing to disclose.

References

[1] A. Turchin et al., Comparison of information content of structured and narrative text data sources on the example of medication intensification, *Journal of the American Medical Informatics Association*. **16** (2009) 362–370.
[2] B. Hazlehurst et al., Natural language processing in the electronic medical record: Assessing clinician adherence to tobacco treatment guidelines, *American Journal of Preventive Medicine*. **29** (2005) 434–439.
[3] Özlem Uzuner et al., Extracting medication information from clinical text, *Journal of the American Medical Informatics Association*. **17** (2010) 514–518. doi:10.1136/jamia.2010.003947.
[4] S. Semler et al., German Medical Informatics Initiative: A National Approach to Integrating Health Data from Patient Care and Medical Research, *Methods of Information in Medicine*. **57** (2018) e50–e56. doi:10/gdzg6p.
[5] H.-U. Prokosch et al., MIRACUM: Medical Informatics in Research and Care in University Medicine: A Large Data Sharing Network to Enhance Translational Research and Medical Care, *Methods of Information in Medicine*. **57** (2018) e82–e91. doi:10/gdzpqv.
[6] S.N. Murphy et al., Architecture of the Open-source Clinical Research Chart from Informatics for Integrating Biology and the Bedside, in: Annual Symposium Proceedings / AMIA Symposium, 2007: p. 5.
[7] G. Hripcsak et al., Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers, *Studies in Health Technology and Informatics*. (2015) 574–578.
[8] J.M. Overhage et al., Validation of a common data model for active safety surveillance research, *Journal of the American Medical Informatics Association*. **19** (2012) 54–60. doi:10/fnbzbb.
[9] A.A. Kimia et al., An introduction to natural language processing: How you can get more from those electronic notes you are generating, *Pediatric Emergency Care*. **31** (2015) 536–541.
[10] Universitätsklinikum Erlangen - Zahlen und Fakten, *Universitätsklinikum Erlangen - Zahlen Und Fakten*. (2020). https://www.uk-erlangen.de/presse/zahlen-und-fakten/.
[11] Averbis health discovery, (2020). https://www.https://averbis.com/de/health-discovery/.
[12] Python 3.6.0, (2020). https://www.python.org/downloads/release/python-360/.
[13] R Core Team, R: A language and environment for statistical computing, (2019). https://www.R-project.org/.
[14] L. Deleger et al., Building Gold Standard Corpora for Medical Natural Language Processing Tasks, (n.d.) 10.