The Importance of Health Informatics in Public Health during a Pandemic
J. Mantas et al. (Eds.)

doi:10.3233/SHTI200502

Use of Natural Language Processing for Precise Retrieval of Key Elements of Health IT Evaluation Studies

Verena DORNAUER [a,1], Franziska JAHN [b], Konrad HOEFFNER [b], Alfred WINTER [b] and Elske AMMENWERTH [a]
[a] *Institute of Medical Informatics, UMIT – Private University for Health Sciences, Medical Informatics and Technology GmbH, Hall in Tirol, Austria*
[b] *Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Germany*

Abstract. Having precise information about health IT evaluation studies is important for evidence-based decisions in medical informatics. In a former feasibility study, we used a faceted search based on ontological modeling of key elements of studies to retrieve precisely described health IT evaluation studies. However, extracting the key elements manually for the modeling of the ontology was time and resource-intensive. We now aimed at applying natural language processing to substitute manual data extraction by automatic data extraction. Four methods (Named Entity Recognition, Bag-of-Words, Term-Frequency-Inverse-Document-Frequency, and Latent Dirichlet Allocation Topic Modeling were applied to 24 health IT evaluation studies. We evaluated which of these methods was best suited for extracting key elements of each study. As gold standard, we used results from manual extraction. As a result, Named Entity Recognition is promising but needs to be adapted to the existing study context. After the adaption, key elements of studies could be collected in a more feasible, time- and resource-saving way.

Keywords. health IT, evaluation studies, natural language processing

1. Introduction

The search and retrieval of health IT evaluation studies via PubMed just by using MeSH terms can be challenging due to lack of precise MeSH terms in the field of health IT evaluation [1]. Having precise information about health IT evaluation studies is important for evidence-based decisions in medical informatics. To solve this issue, in an earlier study, we have used a faceted search based on ontological modeling of key elements of studies to retrieve precisely described health IT evaluation studies [2]. However, retrieving the key elements for the faceted search manually was time and resource-intensive. To solve this issue, we now aimed at applying natural language processing based on the ontology population approach to substitute manual data extraction by retrieving the key elements of the studies by automatic extraction [3]. The

[1] Corresponding Author, Verena Dornauer, Institute of Medical Informatics, UMIT – Private University for Health Sciences, Medical Informatics and Technology GmbH, Hall in Tirol, Austria; Email: verena.dornauer@umit.at.

aim of this paper is to compare four methods for automatic data extraction with regard to their effectiveness.

2. Methods

We compared four natural language processing methods: Bag-of-Words (BoW), Term-Frequency-Inverse-Document-Frequency (TFIDF), Latent Dirichlet Allocation (LDA) Topic Modeling and Named Entity Recognition (NER). We applied each of them to 24 health IT evaluation studies after generating text corpora and pre-processing of the studies. We aimed at extracting the following key elements of health IT evaluation studies: evaluated application system, features of the evaluated application system, organizational unit and users of the evaluated application system, and software product name. Furthermore, we focused on the author of the study, the used study method and the outcome criteria of the study. The programming language Python and the Integrated Development Environment (IDE) PyCharm were used. We utilized the following Python libraries: gensim, NLTK, pandas, pyLDAvis, python-docx, scikit-learn, scispacy and spacy [4]. The detailed information of the experiment and code can be found on the github repository of this study [5]. We compared the effectiveness of the methods by comparing the extracted key elements with the outcome from manual data extraction. The efficiency was measured by observing the coding task due to complexity of the code and the reusability of the Python code.

3. Results

The three statistical-based natural language processing methods Bag-of-Words (BoW), Term-Frequency-Inverse-Document-Frequency (TFIDF) and Latent Dirichlet Allocation (LDA) Topic Modeling were not helpful to extract key elements of the studies. The results of the three methods mostly present single words. Table 1 illustrates one example for each of these 3 methods for the key elements evaluated application system and outcome criteria of the study[2].

Table 1. Examples for the statistical-based natural language processing methods a) Bag-of-Words (BoW), b) Term-Frequency-Inverse-Document-Frequency (TFIDF) and Latent Dirichlet Allocation (LDA) Topic Modeling.

Method	Extracted key element of the study	Results with the natural language processing method	Results with manual data extraction
BoW	Application System	nursing system	nursing information system
	Outcome Criteria	users satisfaction	user interaction satisfaction
TFIDF	Application System	information nursing systems	see above

[2] J. Liaskos, and J. Mantas, Measuring the user acceptance of a Web-based nursing documentation system., *Methods Inf. Med.* **45** (2006) 116–20.

	Outcome Criteria	satisfaction user	see above
LDA TM	Application System	nursing system information	see above
	Outcome Criteria	user satisfaction	see above

The rule-based natural language processing method Named Entity Recognition (NER) showed more promising results. Table 2 illustrates this with one example of the same study.

Table 2. Example result for the rule-based natural language processing method Named Entity Recognition (NER)

Method	The extracted element of the study	Results with the natural language processing method	Result with manual data extraction
NER	Application System	nursing information system	nursing information system
	Outcome Criteria	user interface satisfaction	user interaction satisfaction

Overall, 19 of 24 evaluated application systems, 14 of 15 software product names and 5 of 24 outcome criteria could be extracted correctly with Named Entity Recognition (NER).
All four natural language processing methods were efficiently usable due to high code reusability within each method.

4. Discussion

Named Entity Recognition (NER) is promising for the retrieval of four key elements of health IT evaluation studies: the evaluated application system, the software product name, seldom the outcome criteria of the study and mostly all authors and co-authors. To increase the probability to retrieve also other key elements and so improve the effectiveness of the method, the Named Entity Recognition (NER) method needs to be adapted. Here, so-called entity lists, usable in the Python library spacy [6], could be used. The entity lists could be generated based on the already existing terms of the ontology. By using these entity lists, already known modeled key elements could be detected in a time- and resource-saving task. These entity lists could also serve in the future as a general speech model in the field of medical informatics within several natural language processing applications. Such a general speech model is currently missing in the field of medical informatics. The challenge for our use case is still the unsolved procedure of handling not yet known key elements of health IT evaluation studies. Here, we aim at investigating machine learning methods with parts of active learning, where the person who extracts the key elements of a study can intervene constantly and actively in the machine learning process [7].

The other investigated natural language processing methods Bag-of-Words (BoW), Term-Frequency-Inverse-Document-Frequency (TFIDF) and Latent Dirichlet Allocation (LDA) Topic Modeling as used were not useful for our context. As observed,

they only extract single words, which were not sufficiently similar to the manually extracted key elements. However, Topic Modeling with n-grams could be investigated more due to the possibility to detect more word phrases in this way [8].

Nevertheless, a further limitation of all four methods is the missing context, which is not visible in the results of the analyses, and the reader normally gains while reading an article. Here wrong assumptions on the significance of results to be indeed the right key element are possible. Due to this circumstance, natural language processing methods and here mostly Named Entity Recognition (NER) are only helpful for a semi-automatic data extraction of key elements of studies.

5. Conclusions

The investigated natural language processing methods Bag-of-Words (BoW), Term-Frequency-Inverse-Document-Frequency (TFIDF), Latent Dirichlet Allocation (LDA) Topic Modeling and Named Entity Recognition (NER) cannot substitute the manual extraction of key elements of health IT evaluation studies. However, the Named Entity Recognition (NER) method is promising for semi-automatic data extraction but needs to be adapted. After this adaption, a general speech model for several applications of natural language processing in the field of medical informatics could be gained.

Acknowledgments

This study took place within the project “HITO – a Health IT Ontology”, funded by the Austrian Science Foundation FWF (I 3726-N31) and the German DFG (WI 1605/11-1).

References

[1] Dixon B, Zafar A, McGowan J, Development of a taxonomy for health information technology, Stud. Health Technol. Inform. 129 (2007) 616–620.
[2] Dornauer V, Ghalandari M, Hoeffner K, Jahn F, Winter A, Ammenwerth E, Developing and implementing a health IT ontology for facilitating retrieval of health IT evaluation studies, in: Poster GMDS-Annual Meet. 8. – 11.9.2019, Dortmund., 2019. doi:10.3205/19gmds166.
[3] Lubani M, Noah SAM, Mahmud R, Ontology population: Approaches and design aspects, J. Inf. Sci. 45 (2019) 502–515. doi:10.1177/0165551518801819.
[4] Python Software Foundation, PyPI The Python Package Index, (2019). https://pypi.org/ (accessed October 26, 2019).
[5] Dornauer V, NLPApplicationOnHealthITEvaluationStudies, Github Repos. (2020). https://github.com/VerenaDornauer/NLPApplicationOnHealthITEvaluationStudies (accessed May 11, 2020).
[6] Explosion AI, Linguistic Features spaCy Usage Documentation, 2019. https://spacy.io/usage/linguistic-features#named-entities (accessed November 27, 2019).
[7] Olsson F, A literature survey of active machine learning in the context of natural language processing, Kista, 2009.
[8] Perkins J, Fattohi F, Python 3 text processing with NLTK 3 cookbook : over 80 practical recipes on natural language processing techniques using Python’s NLTK 3.0, 2014.