

Proposal of Semantic Annotation for German Metadata Using Bidirectional Recurrent Neural Networks

Hannes ULRICH^{a,1}, Hristina UZUNOVA^b, Heinz HANDELS^{b,c} and
Josef INGENERF^{a,c}

^a IT Center for Clinical Research (ITCR-L), University of Lübeck, Germany

^b German Research Center for Artificial Intelligence, Lübeck, Germany

^c Institute of Medical Informatics, University of Lübeck, Germany

Abstract. The distributed nature of our digital healthcare and the rapid emergence of new data sources prevents a compelling overview and the joint use of new data. Data integration, e.g., with metadata and semantic annotations, is expected to overcome this challenge. In this paper, we present an approach to predict UMLS codes to given German metadata using recurrent neural networks. The augmentation of the training dataset using the Medical Subject Headings (MeSH), particularly the German translations, also improved the model accuracy. The model demonstrates robust performance with 75% accuracy and aims to show that increasingly sophisticated machine learning tools can already play a significant role in data integration.

Keywords. Metadata, Unified Medical Language System, Deep Learning

1. Introduction

The digital transformation is progressively changing our healthcare system and its disciplines. The rapid pace of digitization is also creating more new clinical data sources that need to be analyzed and integrated to be used together. This is of enormous importance for patient care because a data fusion of all sources allows a comprehensive, holistic overview. The fragmentation of digital healthcare into many proprietary individual systems and data formats makes the desired overview difficult and slows down technical innovations. The clinical data integration shall gap this fragmentation and is an essential foundation for further data processing. One suitable tool in this context is metadata [1], which is able to describe the diverse characteristics of information objects precisely. In addition to content and administrative information, they also suitably depict the structure and - with the help of annotations - the semantics. The semantic coding enables a better understanding of the described data but is only useful if the annotation is carried out extensively and is sustainable. It must be ensured that codes or the coding system also fit the described content of the metadata and that the annotations are carried out consistently [2]. In contrast, inconsistent annotations degrade

¹Corresponding Author, Hannes Ulrich, IT Center for Clinical Research, Lübeck; Telephone: +49 (0) 451 - 3101 5607; Fax: +49 451 3101 5604; E-mail: h.ulrich@uni-luebeck.de

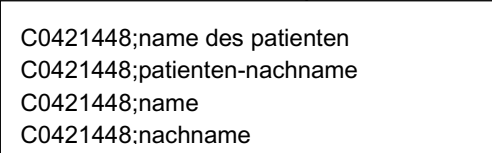
data quality and findability. Manual annotation is time and resource-consuming, so machine support is desirable [3]. However, this support should not simply propose codes but should incorporate previous annotations to ensure consistent annotation. The already annotated datasets provide implicit knowledge about previous annotations that should be used meaningfully. Neural networks are the state of the art to exploit this implicit knowledge and to support annotators. To our knowledge, there are no comparable approaches in the literature that describe a predictive model of semantic annotations for given German metadata.

2. Methods

The proposed approach uses a trained neural network-based model to propose semantic codes corresponding to a given metadata item. In the following, the used dataset and its preprocessing steps are described, as well as the architecture of the neural network.

2.1. Dataset and Preprocessing

The dataset originates from the MDM Portal [4], which collects medical data models, mostly electronic Case Report Forms (eCRFs) from clinical research form data. The MDM currently contains ca. 24.000 forms with 500.000 metadata items in 53 languages. The special characteristic of the MDM is the manual annotations of the forms using Unified Medical Language System (UMLS) codes [5], which is done by medical experts. Since our work focuses on the annotation of German metadata, only forms in German were considered. A total of approx. 150.000 annotated German items were available, which described the question groups, the questions, and structured answer options. Before the dataset was used for training, it had to be cleaned, and then additional data was augmented. The cleaning removed samples with deprecated annotations and codes with less than 50 occurrences were sorted out. To increase the amount of training data and increase robustness, the dataset was augmented using the Medical Subject Headings (MeSH) [6]. The MeSH thesaurus which is mainly used for indexing and retrieval of literature appearing in MEDLINE/PubMed, provides German translations of subject headings. As this German version is included within the UMLS metathesaurus, the selection of codes leads to additional 11.700 samples.



```
C0421448;name des patienten  
C0421448;patienten-nachname  
C0421448;name  
C0421448;nachname
```

Figure 1. The input data are pairs of a UMLS code and the corresponding phrase, for example, the patient surname (in German).

2.2. Network Architecture and Training

The expected input is a metadata definition, a sequence of words representing the display text of the metadata item, e.g., “The name of the patient”. Sequential deep learning models, particularly recurrent neural networks (RNNs) and bidirectional long short-term memories (BiLSTMs), have shown robust performance on tasks which require the

encoding of short sequences of words [7, 8]. In order to predict the most probable semantic annotation, the input sequences are passed through the layers of a LSTM-based network responsible for transforming the definition into a computable representation, learning the sequential correspondences within the words of the definition, and predicting the most probable semantic code. The network architecture is shown in Figure 1. At first, the sequences are split into word tokens, and each unique word is represented through an index. These indices are passed into an embedding layer that transforms each word into a vector representation. These representations are passed into a two-tier BiLSTM layer and then into the last classification layer. The output is mapped to possible classes representing UMLS codes.

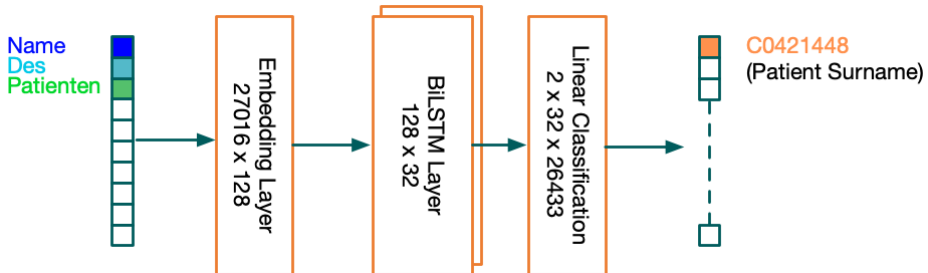


Figure 1. The trained network is built of three consecutive layers. The embedding layers encoded the words into a computable vector representation, which will be used in the next layer. In the two-cell bidirectional LSTM, the word sequences (“Name des Patienten”) are learned and the resulting output is mapped on the labels in the linear layer.

The network was implemented in Python using Pytorch and is available on Github [9]. The embedding layer was parametrized with 26433 unique words and 128 dimensions, the two-tier BiLSTM with 32 hidden dimensions. The input and output dimensions of the linear classifier were set to 25992 representing every possible code in the datasets and with two times of the hidden dimensions of the BiLSTM since the output dimension is double due to the fact of using a BiLSTM.

We use an 80/20 training/test split. 20% of the training data are then used for validation during the training to determine the optimal number of training epochs. An Adam-optimizer with a learning rate of 0.01 is used. The training was carried out for 250 epochs with a batch size of 5000. A cross-entropy loss function is used weighted by a factor $w = l/l_{max}$ where l is the current length of the phrase and l_{max} is the maximal phrase length in the training set. The intuition behind this weighting is to give each word the same relative importance to the loss, such that short and one-word phrases do not overweight long phrases.

3. Results

Different variants were trained to identify the best network configuration. The depth of the BiLSTM layer was varied to achieve the best accuracy on the given data set. The best configuration was trained for another experiment with German GloVe embeddings [10] instead of the inbuilt embedding layer. These embeddings were trained on large text bodies and can better detect synonyms in the data. However, the accuracy was lower compared to the learned embeddings. Then the configured networks were trained with and without the augmented data. The results of the experiments are shown in Table 2,

where T_k indicates that the predicted label was contained in the top k most probable predictions. The best results were achieved by a two-tier BiLSTM closely followed by the one-tier architecture. Furthermore, the results underline the importance of augmented data and its potential to achieve even higher accuracy. Overall, a fair accuracy of ca. 67% could be achieved, where an even higher accuracy of 75% can be observed when the ten best predictions are considered.

Table 1. The table shows the first three predictions for the given set with the probabilities in brackets.

Phrase	1. Prediction	2. Prediction	3. Prediction
Name des Patienten (name of the patient)	Patient surname (25.04%)	Medication name (18.32%)	Patient forename (17.45%)
Krebs der Niere (cancer of the kidney)	Kidney cancer (24.84%)	Subject Diary (17.16%)	Malignant neoplasm of kidney (16.12%)
Krebs der Leber (cancer of the liver)	Malignant Placental Neoplasm (21.20%)	Secondary malignant neoplasm of liver (18.50%)	Liver reconstruction (17.72 %)
Blut (blood)	Blood (19.38%)	Blood in Urine (17.81%)	Coagulation Process (17.21%)

Table 2. The table presents the results of the conducted experiments. We trained models with four different configurations: three different layer depths of the LSTM layer and additionally with pre-trained GloVe embeddings for the best model configuration. The experiments were conducted in each configuration with and without the augmented data. Here, T1, T3, T5, T10 signifies whether the ground truth class was within the first, first three, first five, or first ten most likely predictions. The best results were achieved by the two-tier BiLSTM.

Experiment	T1 Acc.	T3 Acc.	T5 Acc.	T10 Acc.
BiLSTM w/o Augmentation	63.77	70.16	71.39	72.69
BiLSTM w/ Augmentation	67.04	72.66	73.9	75.44
2BiLSTM w/o Augmentation	64.63	70.95	71.2	73.27
2BiLSTM w/ Augmentation	67.27	72.85	74.23	75.63
2BiLSTM w/o Glove w/o Aug.	60.09	66.16	67.34	68.76
2BiLSTM w/ Glove w/o Aug.	65.09	70.56	72	73.64
3BiLSTM w/o Augmentation	63.46	69.17	70.42	71.53
3BiLSTM w/ Augmentation	66.23	71.45	72.66	74.03

4. Discussion

The specificity of the proposed model resides in the processing and semantic annotation of German metadata - to our knowledge, there is no comparable model. The accuracy of the predictions is sound, although there is potential for improvement. In addition, only a fraction of all UMLS Codes were included in the data set, so not all concepts can be predicted. For example, the concept “liver cancer” was not in the dataset, so the network recognized liver and cancer distinctly, but the joining concept was unknown, as seen in [Table 1](#). The use of augmented data showed an improvement in accuracy in all experiments. Therefore, further augmentation sources should be used in subsequent studies to enhance the training dataset. The use of pre-trained GloVe embeddings should allow a better understanding of synonyms. Nevertheless, the overall accuracy was worse. One possible explanation is that most keywords in the phrases are specialized medical vocabularies and are often not included in models pre-trained on general text corpora. Future work will consider and further refine training on the target GloVe embedding dataset. Another possibility is the use of newer attention-based models such as

Transformer, but there is currently no biomedical Transformer model for the German language. One inaccuracy in general remains for the process of data integration using the trained model: the predictions are based on previous annotations. Misclassifications then carry over into subsequent processing steps. However, the MDM portal can also draw on great prior work in terms of consistency and interrater variability [11], so that the use of the model can be recommended.

5. Conclusions

This work is intended as a proof-of-concept to show that increasingly performant machine learning tools can already play an important role in data integration - effectively a stage before the initial provision of curated research data. The proposed model can usefully support annotators to enable new datasets for secondary research and hopes to be an impetus for future work in the area, such as integrating the UMLS graph knowledge into the network.

Acknowledgments

The authors would like to thank the MDM team for their long-standing and great work of providing and annotating medical forms. This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) DFG grants IN 50/3-2.

References

- [1] Ulrich H, Kock-Schoppenhauer AK, Deppenwiese N, Gött R, Kern J, Lablans M, Majeed RW, Stöhr MR, Stausberg J, Varghese J, Dugas M, Ingenerf J. Understanding the Nature of Metadata: Systematic Review. *J Med Internet Res.* 2022 Jan 11;24(1):e25440. doi: 10.2196/25440. PMID: 35014967; PMCID: PMC8790684.
- [2] Cardoso SD, Pruski C, Da Silveira M, Lin Y-C, Groß A, Rahm E, et al., Leveraging the impact of ontology evolution on semantic annotations, in: Springer, 2016: pp. 68–82.
- [3] Ulrich H, Kock-Schoppenhauer AK, Andersen B, Ingenerf J. Analysis of Annotated Data Models for Improving Data Quality. *Stud Health Technol Inform.* 2017;243:190-194. PMID: 28883198..
- [4] Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, Varghese J. Portal of medical data models: information infrastructure for medical research and healthcare. *Database (Oxford).* 2016 Feb 11;2016:bav121. doi: 10.1093/database/bav121. PMID: 26868052; PMCID: PMC4750548.
- [5] Bodenreider O, The unified medical language system (UMLS): integrating biomedical terminology, *Nucleic Acids Research.* 32 (2004) D267–D270.
- [6] ZB MED – Information Centre for Life Sciences, German MeSH, ZB MED - Informationszentrum Lebenswissenschaften. (2022). <https://www.zbmed.de/en/open-science/terminologies/german-mesh/> (accessed January 21, 2022).
- [7] Xu K, Zhou Z, Hao T, and Liu W, A bidirectional LSTM and conditional random fields approach to medical named entity recognition, in: Springer, 2017: pp. 355–365.
- [8] de Buy Wenniger GM, van Dongen T, Aedmaa E, Kruitbosch HT, Valentijn EA, and Schomaker L, Structure-tags improve text classification for scholarly document quality prediction, in: 2020: pp. 158–167.
- [9] Ulrich H, mentored, Zenodo, 2022. doi:10.5281/zenodo.5897287.
- [10] Pennington J, Socher R, and Manning CD, Glove: Global vectors for word representation, in: 2014: pp. 1532–1543.
- [11] Varghese J, Sandmann S, Dugas M. Web-Based Information Infrastructure Increases the Interrater Reliability of Medical Coders: Quasi-Experimental Study. *J Med Internet Res.* 2018 Oct 15;20(10):e274. doi: 10.2196/jmir.9644. PMID: 30322834; PMCID: PMC6231825.