Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200127

Character-Level Neural Language Modelling in the Clinical Domain

Markus KREUZTHALER [a,b,1], Michel OLEYNIK [a] and Stefan SCHULZ [a]
[a] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz*
[b] *CBmed GmbH — Center for Biomarker Research in Medicine*

Abstract. Word embeddings have become the predominant representation scheme on a token-level for various clinical natural language processing (NLP) tasks. More recently, character-level neural language models, exploiting recurrent neural networks, have again received attention, because they achieved similar performance against various NLP benchmarks. We investigated to what extent character-based language models can be applied to the clinical domain and whether they are able to capture reasonable lexical semantics using this maximally fine-grained representation scheme. We trained a long short-term memory network on an excerpt from a table of de-identified 50-character long problem list entries in German, each of which assigned to an ICD-10 code. We modelled the task as a time series of one-hot encoded single character inputs. After the training phase we accessed the top 10 most similar character-induced word embeddings related to a clinical concept via a nearest neighbour search and evaluated the expected interconnected semantics. Results showed that traceable semantics were captured on a syntactic level above single characters, addressing the idiosyncratic nature of clinical language. The results support recent work on general language modelling that raised the question whether token-based representation schemes are still necessary for specific NLP tasks.

Keywords. Neural Networks, Electronic Health Records, Natural Language Processing

1. Introduction and Motivation

So-called word embeddings have become popular for various natural language processing (NLP) tasks. They capture the semantics of a word to a certain degree and have replaced term-weighted vector-based representation schemes from the pre-deep learning era. For building such embeddings, a critical amount of textual data is needed to obtain stable vector representations with different surrounding contexts. The excellent performance of such representations, in combination with different neural network architectures were demonstrated in various clinical NLP tasks.

In most cases, character-based input representations were handled as a merely supportive information layer for word-level embeddings in different NLP problem domains.

[1]Corresponding Author: Institute for Medical Informatics, Statistics and Documentation, Graz General Hospital and University Clinics, Auenbruggerplatz 2, 8036 Graz, Austria, Europe; CBmed GmbH — Center for Biomarker Research in Medicine; E-mail: markus.kreuzthaler@medunigraz.at.

The idea that character-level neural language models (CNLM) alone, based on recurrent neural networks (RNN) and more specifically long short-term memories (LSTM) are able to sufficiently capture the information needed for a given task recently got again attention by various research groups. The main motivation behind these experiments — Google named them "tokenizer-free language models" [1] — is to address NLP tasks using only single characters as input representations, thus capturing human language on its most granular representation level [2].

In this work we investigated the possibility of character-based *clinical* language modelling in order to address its well-known idiosyncrasies (telegram-style, single-word composition, spelling variants, and typing errors) with the purpose of optimally capturing the meaning of real-world textual data.

The paper is structured as follows: Section 2 gives an overview of recent work on CNLMs. Section 3 describes the clinical data set for the investigation as well as used methods and language models. Section 4 exemplifies the results, gives insight into the lexical semantics of the obtained embeddings and discusses the results. Finally, Section 5 concludes the paper and provides an outlook for further investigation.

2. Background and Related Work

Zhang and LeCun [3] applied convolutional neural networks (CNN) on a single character input level for ontology classification, sentiment analysis and text categorization. They stacked six convolutional layers and three fully-connected layers for the specific tasks. The experimental design reached a test set based accuracy of 0.98 for ontology classification (DBpedia), 0.95 for sentiment analysis using the Amazon review polarity data set, and 0.60 exploiting the Amazon review full score data set. An accuracy of 0.71 was reached for the Yahoo! Answers topic classification data set.

Kim et al. [4] applied single character-level inputs to a deep neural network model consisting of a CNN, a highway network over characters, and an LSTM at the end. They tested their model on the Penn Treebank for different languages and achieved state-of-the-art results with less parameters in comparison to the other machine learning based approaches. They reported a perplexity of 78.9 on the English Penn Treebank test set. Their character-level input models outperformed word/morpheme-based input representation schemes in general. The model was also shown to be able to capture different levels of semantics on a token level constituted by a time-series based input representation scheme of single characters.

More recently, Choe et al. [1] from Google showed that tokenizer-free CNLMs are competitive with current state-of-the-art language models using the One Billion Word benchmark in combination with a deep transformer neural network containing 40 self-attention layers. By using several semantic benchmarks, they showed the ability of their network to capture layer-wise word-level semantics even though the input representation was character-based. This work corroborates another recent work from Stanford and Facebook [5], which concluded that CNLMs based on recurrent neural networks (e.g. LSTMs) are able to capture upper-level semantics.

Motivated by the research mentioned above, we explored to what degree CNLMs can capture token-level lexical semantics from clinical texts. We accomplished that by fetching embeddings out of an RNN, after training it on a critical mass of real-world data,

viz. ICD-10 coded problem list entries from a clinical information system, modelled as a time series of one-hot encoded single characters. With a nearest neighbour search on selected clinical narrative content we fetched word types within the embedding space and hypothesized that they were semantically connected to the given input. To the best of our knowledge, this is the first investigation on the extent of CNLM-based lexical semantics that can be obtained from clinical narratives.

3. Methods and Data

The initial data set is composed of ∼1.9 million unique 50 characters long problem list descriptions in German, annotated with ICD codes. These annotations had been done by physicians on in-patient discharge, the ICD codes being relevant for billing. To have a sufficient number of initial samples per class we took the top-100 most frequent three-character ICD-10 codes from a 90% split and up-sampled them to the most common one to balance the whole data set. The final data consisted of ∼6 million coded problem list entries used for CNLM training downstreamed to their corresponding disease code.

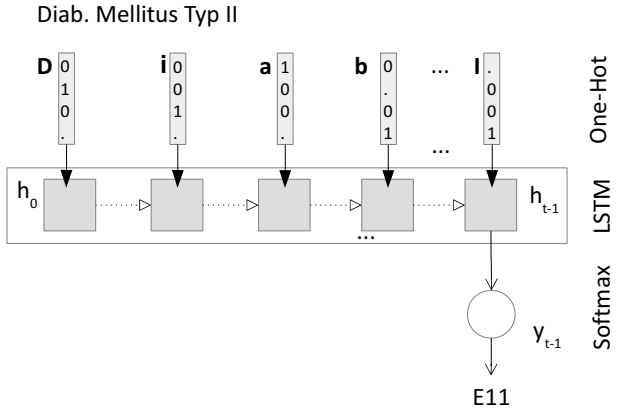

Figure 1. Schematic description of the LSTM network in use [6].

After having prepared the training data set, we fed it to an LSTM network as depicted in Figure 1. The problem list entry is represented as a series of one-hot encoded input vectors of dimensionality 122 (number of alphabetic characters in the training corpus) to the recurrent neural network. We refrained from any pre-processing and used all strings as they were in the raw data. The neural network was trained for 5 epochs (∼73 hours) using the deeplearning4j[2] framework on an NVIDIA Titan V GPU with the following parameters: mini-batch size: 512; learning rate: 1.0; L2 regularisation: 1E-6; weight initialisation: XAVIER; updater: RmsProp; LSTM hidden state vector size: 128; number of LSTM cells: 60; LSTM activation: TANH; output layer activation: SOFTMAX; output layer loss function: MCXENT; total number of network parameters: 141,412, which is about a factor of 40 less than available training observations.

[2]https://deeplearning4j.org/

In order to obtain a specific character-induced word embedding, we accessed the hidden state h_{t-1} out of the cell at step $t-1$ for a given token $\mathbf{X}$ of length t. Using the k-d tree data structure we built a searchable $k = 128$ dimensional index for all occurring types in the training corpus to get back the *top n* nearest neighbours given an input vector. The input vector was inferred the same way using the trained LSTM network as described before. With the trained LSTM network and the k-d tree data structure, we could obtain the nearest neighbours in reasonable time by exploiting the cosine similarity measure in the embedding space. This space was used to test our hypothesis that nearest neighbours to a given input vector contain verifiable semantics with respect to the initial search vector.

4. Results and Discussion

Table 1 lists the top 10 nearest neighbours to a given concept in descending order. The first column contains the results for the disorder "Nikotinabusus" (nicotine abuse). Interestingly, most of the nearest embeddings to the given token contain variations of the clinical quantification of smoking cigarettes in pack years (py), a level of detail that is irrelevant for ICD. The pack year (py) is the unit in which the inhaled smoke dose of a cigarette smoker is approximately described. Moreover, the quantitative values correspond to the manifestation of a harmful use of nicotine via some smoking behaviour.

"Nikotinabusus"		*"Handverletzung"*	
Word type	Sim.	Word type	Sim.
chronisch	1.00	Handverletzung	1.00
(15/die)	0.99	Strecksehnendurchtr	0.91
abuse)	0.99	Fingernagelbruch	0.81
(20Stk./d)	0.99	Zeigefingergrundgliedfraktur	0.80
(25/d)	0.99	ausw.kons.	0.80
Z/d)	0.99	Handwurzelkontusion	0.79
(135	0.99	eingestauchte	0.79
(Abstin.seit	0.98	Daumenendphalanxstauchungsfraktur	0.78
(30pys)	0.98	Handwurzelknochen-Fraktur	0.78
(50pj)	0.98	4.links	0.77

Table 1. Top 10 nearest neighbors with corresponding similarity scores for two example words.

The second column contains the most similar character-level-induced word embeddings for "Handverletzung" (hand injury). The majority of the top 10 results contain different types of injuries of the body structure hand, and anatomical locations which are part of the hand. A remarkable result is the semantic response "Daumenendphalanxstauchungsfraktur" (thumb end phalanx compression fracture). Semantically obviously related to each other, the lexical representation of "Handverletzung" and "Daumenendphalanxstauchungsfraktur" has nearly nothing in common with a Levenshtein distance of 25.

5. Conclusion and Outlook

In this paper we investigated to what extent a CNLM with an LSTM network architecture is able to induce meaningful lexical semantics in the clinical domain. We trained the network on ∼6 million ICD-10 coded problem lists, a critical amount of training data for language modelling. After the training, we fetched the embeddings by accessing the hidden state h_{t-1} within the cell at the last step of an input word of the LSTM network. We then indexed the resulting 128-dimensional character-level-induced word embeddings using a k-d tree data structure and, via a top 10 nearest neighbour search, obtained the most similar whitespace separated tokens given an input word. We showed that meaningful lexical semantics could be captured at this representation level using an exploratory evaluation.

Our results support recent investigations that single character-level representations to RNNs are able to fetch semantics in contrast to the most low-level fine-granular input scheme. The work presented here uses a single LSTM network in contrast to more complex models from recent research on deep transformer models like BERT [7] or stacked Bi-LSTMs like ELMo [8] for contextual embedding modelling.

Furthermore, this preliminary evaluation was done purely on a human interpretation level on selected examples. Future investigations of CNLMs in the clinical domain should consider objective evaluation measurements for semantic similarity and language model quality. Besides the mentioned limitations, the work presented in this paper supports the promising potential of CNLMs in processing raw clinical language in a robust but nevertheless semantically precise manner.

References

[1] Dokook Choe, Rami Al-Rfou, Mandy Guo, Heeyoung Lee, and Noah Constant. Bridging the gap for tokenizer-free language models. *arXiv preprint arXiv:1908.10322*, 2019.

[2] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[3] Xiang Zhang and Yann LeCun. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*, 2015.

[4] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[5] Michael Hahn and Marco Baroni. Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text. *arXiv preprint arXiv:1906.07285*, 2019.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[8] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.