Effectiveness of neural language models for word prediction of textual mammography reports

Mihai David Marin
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
m.marin@student.utwente.nl

ABSTRACT

Radiologists are required to write free paper paper text reports for breast screenings in order to assign cancer diagnoses in a later step. The current procedure requires a lot of time and needs efficiency. To streamline the writing process and keep up with the specific vocabulary, a word prediction tool using neural language models was developed. Challenges as different languages (English,Dutch), small data sizes and low computational power have been overcome by introducing EnDuRLM process, able to improve by 25% the current workflow according to RPE measurement. After defining model architectures, EnDuRLM process involves data preparation, hyperparameters optimization, configuration transfer and evaluation. This work supports future research involving other languages and also an extensive set of real-world applications.

Keywords

mammography, medical reports, neural language model, text generation, natural language processing

1. INTRODUCTION

Breast cancer is the most commonly occurring cancer in women and the second most common cancer overall. There were over 2 million new cases in 2018 globally, with The Netherlands being the third in the top 25 countries with the highest rates of breast cancer [4]. Early-stage breast cancer detection could reduce breast cancer death rates significantly in the long-term. Different screening techniques can be used to diagnose abnormalities, that can indicate cancer [3], e.g. Mammograms and Computerized Tomography (CT) that uses x-rays of distinct wavelengths, Magnetic Resonance Imaging (MRI) which uses magnetic energy and Ultrasound that uses the sound waves etc. Screening mammography has been shown to reduce breast cancer mortality by 38-48% among participants [5].

After applying these imaging techniques by experts (e.g. radiologists), the findings are communicated to the referring doctor in a physical form and meanwhile also digital. To understand the shift to electronic medical records and radiology data information systems, the increased extension of Natural Language Processing (NLP) techniques in health care in past years allowed clinical applications, as information retrieval [13], reports structuring [20] or diagnosis classification [23]. Such tools improve the speed of the process, the accuracy of the diagnosis and in the same time, reduce the number of doctors needed to achieve this task.

A screening report is the key component of breast cancer diagnostic process. A study revealed that each American radiologist interprets on average 1777 mammograms per year, resulting in approximately one new mammogram each working hour [24]. In this paper, we want to streamline the process of unstructured report writing by introducing a word prediction tool, based on neural language modelling. Previous studies shown that word prediction increased the text composition time by at least 22% in long term use [15].

The main contributions of this paper are:

1. The process *English Dutch Radiology Language Modelling (EnDuRLM)*, developed to overcome challenges as different languages, optimization difficulty, computational restrictions and limited corpus size. This approach involves collecting and preprocessing two data sets (English and Dutch) for language modelling, followed by hyperparameters optimization on the English dataset with basic LSTM architecture [22]. Then, the configuration is transferred to the Dutch dataset and the other model architectures: AWD [17] and FRAGE [8]. In the end the models are evaluated using perplexity and best models are selected for further analysis.
2. The metric *Radiology Process Evaluation (RPE)*, created to evaluate the models by measuring their efficiency in the process of cancer diagnosis.

Using language models developed in EnDuRLM and evaluated with RPE allows the development of a vast range of real world applications. The focus is on next word suggestion, where the model predicts the upcoming word based on context provided by previous words. Further applications include: missing data estimation, where lost data can be generated based on context; quality check, where radiologists receive suggestions about grammatical or spelling errors; educational training, where students or residents learn how to write vocabulary and structure specific.

This paper is structured as follows: first, we will present related work. Second, we will describe the datasets we have used. Then, we will explain EnDuRLM method in detail. Finally, we will present the results, draw conclusions and discuss future work. The trained models, the tool and all other code used will be published at Github[1] after publication.

[1]https://github.com/mihaimdm22

31th Twente Student Conference on IT July 5th, 2019, Enschede, The Netherlands.

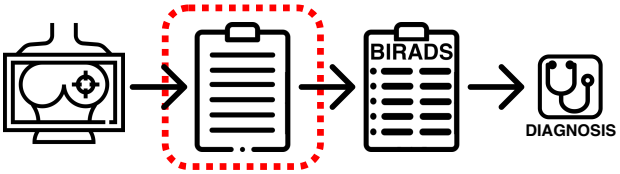


Figure 1. General flow of information for cancer diagnosis using breast imaging.

2. RELATED WORK

In this section, we will discuss automation initiatives for the process of cancer diagnosis using breast imaging, followed by an overview of the language models architectures used.

2.1 Breast imaging automations

Because radiologists interpret so many mammograms and because the proper interpretation of a screening mammogram is often a matter of life or death for the woman involved, various attempts have been made to streamline the mammography reporting process and introduce consistent structure and terminology into mammography reports. The main standard for breast cancer radiology reporting is *"Breast Imaging-Reporting And Data System"* (BI-RADS) [1]. The BI-RADS lexicon provides specific terms to be used to describe findings. Along with that, it also describes the desired report structure, for example, a report should contain breast composition and a clear description of findings[20].

In the process of setting a diagnosis based on a breast image involves a report that describes the observations on the images. Those reports are usually unstructured and NLP-based postprocessing can be used to obtain a structured report [20, 19]. In the end, a doctor sets the diagnosis based on the unstructured report or structured report if available. The complete process is illustrated in Figure 1.

In the last decade Computer Aided Diagnosis solutions [2] as 'SecondLook' (made by iCAD) were developed to help radiologists in reading mammograms. Their efficiency appears to be contradictory because of limited improvements on a long period [14] and therefore their accuracy should be improved to be ultimately considered useful. These limitations have been solved by development of deep learning solutions, which have been shown to achieve near-human performances for some applications [11]. If previous methods rely on regions of interest (parts of image), Zhu et al. [26] proposed an end-to-end approach based on the whole mammogram. The results were modest compared to classic methods, still restricted for a real-world implementation.

In 2009, an algorithm which is capable of assigning BI-RADS final assessment categories from English radiology reports using Natural Language Processing with a precision of approximately 97% accuracy for correct identification [23].

Later on, in 2013, an improvement of language models for radiology speech recognition using n-gram and word frequency in unstructured reports dictation has been developed[21].

More recently, in 2018, using a Conditional Random Field (CRF) model, Pathak et al.[20, 19] developed an algorithm which structures free-text dutch radiology reports on breast cancer for quality assurance. The results close to doctor accuracy (approximately 95%) allowed clinical implementation as annotation tool (TWENTnotator) where an unstructured report turns into Birads structured report automatically.

2.2 Language Models

Because of limited research in radiology combined with natural language processing, we had implemented the basic LSTM as described in PyTorch documentation, inspired from Sherstinsky's paper [22]. Then, we make use of two state of the art architectures retrieved from a repository that tracks the progress in Natural Language Processing (NLPProgress) to reach maximum results.

On top of the architecture described above, Merity et al. proposed a strategy to regularize and optimize the model, outperforming existing approaches [17]. As displayed by its naming (AWD-LSTM), the study introduces a non-monotonically triggered version of the averaged stochastic gradient method (AvSGD) and weight-dropped (WD) LSTM regularization (DropConnect on hidden weights).

The state of the art method offered by Gong et al. [8] makes use of a way to learn frequency-agnostic word embedding (FRAGE) using adversarial training. This representation technique is build at it's own on top of a joint improvement of other's work: AWD-LSTM-MoS [25] and dynamic evaluation [12].

3. DATA SETS

This section describes the original data sets, the steps of preprocessing and feature generation. To provide a fair comparison between English and Dutch language models, we have to input similar datasets regarding content, size and structure.

We decided to choose the following data sets because Dutch is of interest for ZGT and English is available for comparison:

1. ZGT (Hospital Group Twente) Mammography Reports Database in Dutch Language. This database provides approx 48000 reports dated from 2012 to 2017.

2. MIMIC-III (Medical Information Mart for Intensive Care) developed by MIT Lab for Computational Physiology [9] in English Language. This database contains information linked to 53,423 distinct hospital admissions for adult patients (>16 years old) admitted to critical care units of Boston Hospital (Massachusetts, U.S.A) between 2001 and 2012. We will focus on the NoteEvents table specifically, because it contains free text reports. We will further refer to this dataset as MIMIC.

Both datasets were subject to de-identification by removing critical patients data such as id, name, address, date of birth etc. and are provided under a data use agreement.

3.1 Word Embeddings Overview

According to similar studies [6], a qualitative assesment of the word embeddings obtained by training a Continuous Bag Of Words model on the MIMIC dataset has been made. Main objective is to have words with similar context occupy close spatial positions, thus checking if the nearest neighbours of a specific word are semantically similar . A visualization of the ebeidn;dings can be found

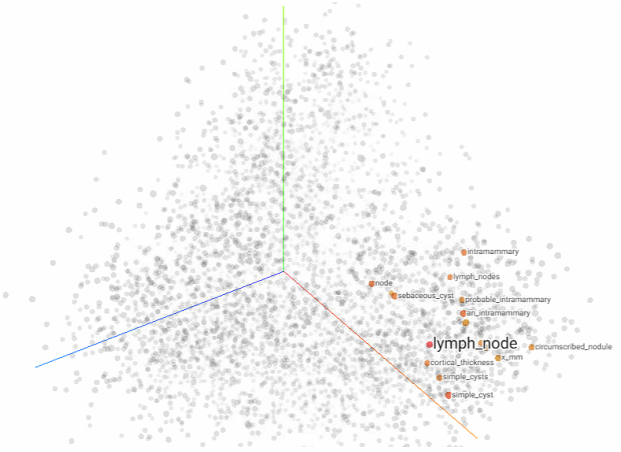


Figure 2. Visualization of the embedding space for MIMIC corpus learnt by a Word2Vec model, highlighting nearest neighbours of the word 'lymph'.

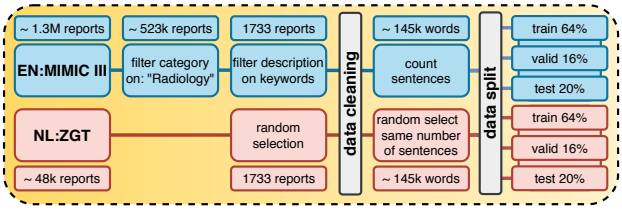


Figure 3. Flow of data in preprocessing step of EnDuRLM

in Figure 2. The embeddings were visualized by means of Principal Component Analysis (PCA), using the top three principal components to reduce the dimensionality of the dataset to three dimensions. As described by Mikolov et al. [18] learnt word embeddings encode many linguistic regularities and patterns that can be represented as linear translations. Furthermore, computing the nearest neighbour words reveals the semantic similarity between neighbouring word vectors. For example, the five nearest neighbours of "lymph" are: "cyst", "circumsied nodule", "hipoechoic lymph", "fibroadenoma", "fluid collection".

3.2 Data preparation

Preprocessing the text is a delicate action which requires choosing the optimal tools given both the data and language models [7]. A typical approach for data preparation is as follows: each of the sets has to be textually reviewed first, in order to apply dataset cleaning of irrelevant data as headings, punctuation, strange names and quoted dialog sequences. The next step is lower casing all the words, so that tokenization can be done. Tokenization is the process of segmenting running text into words and sentences. Figure 3 illustrates clearly step by step the process.

3.2.1 English-Dutch data alignment

The alignment of data is crucial in finding a relation between two languages regarding modelling. Taking into account the differences of the datasets structure and morphology, we decided to have the same number of sentences in both datasets.

Table 1. Aligned examples of report descriptions for datasets

English (MIMIC)	Dutch (ZGT)
Dig diagnostic mammo bilateral	Echo mamma beiderzijds
R mammography specimen right	Mammografie rechts
Mammo needle localization left	MRI mamma punctie links

In order to have the same type of reports, MIMIC's main reports table was filtered on category 'Radiology' resulting in less than half of the size, 523000 reports. The description of the reports varied from brain scan to left foot bone x-ray. A list with all the descriptions was sent to the hospital that provided the ZGT dataset, in order to filter and synchronize with their dataset descriptions as in Table 1. Filtering again on these description resulted in almost 2 thousand reports, less than 5% compared to ZGT reports. In order to have the same count again, we randomly selected the same amount of reports from ZGT database. Then the data cleaning was done individually.

Given the differences in report template and structure, the inequality problem did not dissolve. We counted the sentences and ensured to have the same number of sentences, therefore approximately same number of words: almost 150 thousand.

3.2.2 Data cleaning

Given the differences in layout and structure, the data cleaning was done separately, with the same end goal in mind: lower cased simple sentences without header, numbers or punctuation:

- MIMIC main header (before FINAL REPORT) was removed because it is a computer generated part of the file and thus irrelevant.
- MIMIC personal data or findings are shifted and marked between '[**' and '**]'. Those were replaced by tag '<unk>' (for unknown).
- ZGT reports have a specific structure, where point is replaced by comma. Those were replaced back to point accordingly.
- Measurements values, date and time were also replaced by tag '<unk>'.
- Modify common typos and expand common abbreviations (for e.g. y.o. to years old or dr. to doctor).
- Delete auto generated headers.
- Remove double spaces and tabs.
- Split in sentences using Spacy.
- All sentences were converted to lower case.
- All punctuation was removed.
- Numbers and roman numerals were replaced by N.
- End of each line was replaced by <eos>.

The decision to keep stop words is because on its relevance in final model. In Appendix A, a table showcases an example of each dataset before and after data cleaning.

3.2.3 Data preprocessing

The file containing same number of sentences of cleaned text (retrieved randomly) was subject to split. The dataset was split 80% train and 20% test. The train part was at its time split again in 80% train and 20% validation. These decisions were taken in accordance with [16].

Taking into account the small datasets size, we decided to include all the words in vocabulary. The sizes of the

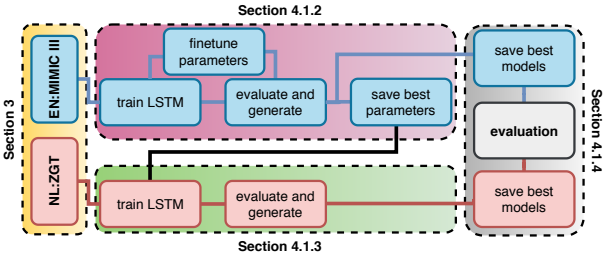


Figure 4. Flow diagram of EnDuRLM process

vocabularies were: 3148 for MIMIC and 4443 for ZGT.

4. METHOD

This section describes our systematic approach to solve the given problem (Section 4.1) and it's real world implementation scenario(Section 4.2).

4.1 EnDuRLM

The English Dutch Radiology Language Modelling (EnDuRLM) framework is designed to overcome challenges as different languages (English,Dutch), small data sizes and low computational power and provides a structural way to obtain good radiology language models. After the data is preprocessed as in Section 3, the hyperparameters optimization is done on the LSTM architecture and MIMIC data set and later on the best configuration is transferred to the other models (AWD and FRAGE) and languages (Dutch: ZGT). In the end evaluation of models by comparison is done.

For a better overview the EnDuRLM process is illustrated in Figure 4, further details following in the next sections.

4.1.1 Model architecture

The LSTM architecture is designed to be better at storing information and finding and learning long-term dependencies than standard recurrent networks. Recent research has shown that a well tuned LSTM baseline can outperform more complex architectures in the task of word-level language modeling.

For this approach we make use of three model architectures: LSTM [22], AWD LSTM [17] and FRAGE LSTM [8]. The code for these implementations is retrieved from open source repositories and adjusted to fit EnDuRLM code.

According to previous studies [16], we used a batch size of 20 and unrolled the network for 35 time steps.

4.1.2 Hyperparameters optimization

All networks were trained with SGD. Merity et al. [17] pointed out that between SGD, Adam, Adagrad and RMSProp, SGD provides better performances. We evaluated our models using the average per-word perplexity on a validation set during training and on a test set after training. We terminated the training process when the validation perplexity had stopped improving for five epochs, and kept the model with the best validation perplexity. Moreover, following initial tests, we decided to stop the training after 10 epochs if validation perplexity was more than 100. All models were trained for a maximum number of 100 epochs. We performed two rounds of random search. In the first round we varied the following parameters between ranges specified in previous work of Merity et al. [17], showcased in Table 2. The second round consisted of restricting the ranges based on the top 10% models regarding values of perplexity from the first random search. The best configuration is saved and used for the other model architectures in the configuration transfer phase.

4.1.3 Configuration transfer

Given the best hyperparameters setting, we apply them on the remaining models for MIMIC dataset and on all models for ZGT dataset. Taking into account that AWD and FRAGE models contain additional hyperparameters for the added features, for this study we will keep their default (dropout for RNN layers: 0.3, dropout for input embedding layers: 0.65, dropout to remove words from embedding layer: 0.1, amount of weight dropout to apply to the RNN hidden to hidden matrix: 0.5, alpha L2 regularization on RNN activation: 2, beta slowness regularization applied on RNN activation: 1, weight decay applied to all weights: $1.2 \cdot e^{-6}$) for a similar dataset, as described in their original papers [17, 8].

4.1.4 Intrinsic Evaluation

Intrinsic evaluation metrics allow to measure the quality of a model independent of a particular application [10]. Most typical metric used to measure the efficiency of a language model is perplexity. As evident in the last line of equation 1, the perplexity is low if the conditional probability of the word sequence is high. Therefore, minimizing the perplexity of a test set is equivalent to maximizing the probability of the test set according to the language model.

The worst model would have a perplexity equal to the size of vocabulary size, because, on average, for each word in the sequence of the data, we have the option to choose any word from the vocabulary. Lowering the perplexity would narrow our options, therefore a better model.

$$\begin{aligned} PP(W) &= P(w_1, \ldots, w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{P(w_1, \ldots, w_N)}} \\ &= \sqrt[N]{\prod_i^N \frac{1}{P(w_i|w_1, \ldots, w_{i-1})}} \end{aligned} \tag{1}$$

where: $w_1, w_2, ..., w_N$ are the words from the test set W with length N

4.2 Real world implementation

Taking into account that consistency is a key feature in report writing, a tool that suggests the next word, based on the previous words (similar to mobile phone keyboard) would streamline the process. The best model for English/Dutch (separately) will be implemented as a feature for the previous version of TWENTnotator[2], a tool developed by University of Twente for ZGT Hengelo for manual/automatic annotation of unstructured reports. The web application has a managerial system for users, standards, reports and projects, in order to handle the whole process of conversion from unstructured to structured reports. A mock-up of the feature is illustrated in Appendix B.

4.2.1 Extrinsic evaluation

Extrinsic evaluation refers to integration of the language model in an application and measuring how much the ap-

[2]https://github.com/yannislinardos/annotationTool

Table 2. Hyperparameters ranges

Parameter	Step size*	Range random search 1	Range random search 2	Range reduction percentage
Embedding size	10	100-800	500-700	71.42%
Number hidden neurons	100	100-2000	1000-1500	73.68%
Dropout probability	0.05	0.10-0.98	0.50-0.85	60.22%
Learning rate	5	5-100	10-30	78.94%
Gradient clipping norm	0.01	0.01-0.80	0.05-0.45	50%
Total				**67.2%**

*the step size is the same for both searches

Table 3. Validation and test perplexities for both datasets

Dataset	Model	Epochs	Valid PPL	Test PPL
MIMIC	LSTM	45	14.08	13.47
	AWD LSTM	39	11.15	10.79
	FRAGE LSTM	42	**9.87**	**9.76**
ZGT	LSTM	56	28.15	27.22
	AWD LSTM	43	**15.45**	**14.94**

plication improves [10]. Implementation of the language models as word prediction in TWENTnotator will facilitate extrinsic evaluation. Moreover, samples of the generated text will be sent to the hospital, so that specialists will evaluate the correctness and relevance of the text.

5. EXPERIMENTS AND RESULTS

Because of computational resources limitations of ZGT (no graphical card computation power), we made the hyperparameters optimization using random search on MIMIC dataset. In the first round of random search, we trained and tested 1000 different LSTM models with the parameter ranges defined in the Table 2. Figure 5 illustrates all the models parameters in parallel, highlighting with colours best 100 models based on test perplexity. This highlighting helped on constraining the ranges for the second round of random search. The reason for a second search lays in the need for robustness and accuracy. This time, the step sizes remained the same, while ranges were reduced by an average of 67.2% accordingly to Table 2. This round of random search trained and tested 100 different LSTM models and the final best performing model, considering the validation set perplexity, had the following configuration:

- Hidden neurons: 1400
- Embedding size: 660
- Learning rate: 30
- Dropout: 0.76
- Gradient clipping norm: 0.28

This configuration was applied to the other model architectures (AWD and FRAGE) for both datasets. The models are evaluated using perplexity metric and the results are shown in Table 3. Comparing the results for MIMIC dataset, we observe an improvement of perplexity in validation set of almost 35% from LSTM to FRAGE which is quite good. The value of perplexities is good, taking into account the small size of the dataset. On the other side, ZGT dataset has an improvement of 65% from LSTM to AWD, and we cannot provide an exact result with FRAGE. The problems encountered when trying to implement FRAGE were technical related, because of the absence of a graphical power unit capable of running CUDA

Table 4. Radiologists evaluation

	Original		Generated		
Radiologist	MIMIC(3)	ZGT(3)	MIMIC(2)	ZGT(2)	Overall accuracy %
1	2	2	2	2	80
2	3	2	2	2	90
3	1	2	2	2	70
4	3	3	2	2	**100**
5	2	3	2	2	90
Total					86

between () is displayed the number of sentences

environment. We can estimate the value of FRAGE for ZGT by aligning with the results from MIMIC of 12% improvement from AWD to FRAGE. The differences between English and Dutch are normal given their disparity and incapability to implement FRAGE for ZGT data set.

To display and compare the models accuracy, we plotted the validation perplexity of the first 50 epochs for each model arhitecture (LSTM, AWD, FRAGE) and each data set(MIMIC, ZGT), resulting in 5 models, because ZGT does not have a FRAGE model. In order to make a clear illustration of how the models fit during training, we decided to display just the first 15 epochs and perplexity of maximum 100. Figure 6 shows high convergence for each setup. The perplexity decrease under 100 in the first 3 epochs and stabilize quite fast for each setup, excluding the LSTM model with ZGT, where the perplexity converges slower.

We analyzed five sentences generated with the best setup and the results are surprising because they seem to make sense and the order of the words is very good. Further analysis using specialized tools will be made in order to assign a statement.

In accordance with the best architecture of each language (FRAGE for English, AWD for Dutch), we created an evaluation form for ZGT hospital. The process involves the expertise of the radiologists, because of their extended knowledge about specific vocabulary and topic. Moreover, they will benefit the real world implementation in the end. Their task involves sentences that could have been generated (entirely) or retrieved from an original report (entirely). The process implies an estimation based on Likert scale whether the sentence was generated by the language models or not (e.g. Strongly Disagree would mean retrieved from original report, Neutral would mean unsure and Strongly Agree would mean generated text). Besides choosing one of the statements of agreement, they have to motivate the answer or give short explanation about decisive points in the evaluation. The sentences are chosen by random and do not relate to each other. The tables with sentences for this part of evaluation are attached to Appendix D. With this evaluation technique, we want to check whether they can distinguish between real report and generated report and see their motivation for this decision. Their arguments can help us further improve the model and maybe find out some similarities.

We have received feedback from five MRON radiologists, three of them specialized in the field of mammography diagnostics and the other two specialized in other fields. For a good overview we decided to count the number of correct guesses and illustrate the results in Table 4. All radiologists correctly guessed the generated sentences for both languages with 100% accuracy, arguing that the sentences make a lot of sense, but there are words that suggest different context. The point where the overall accuracy went

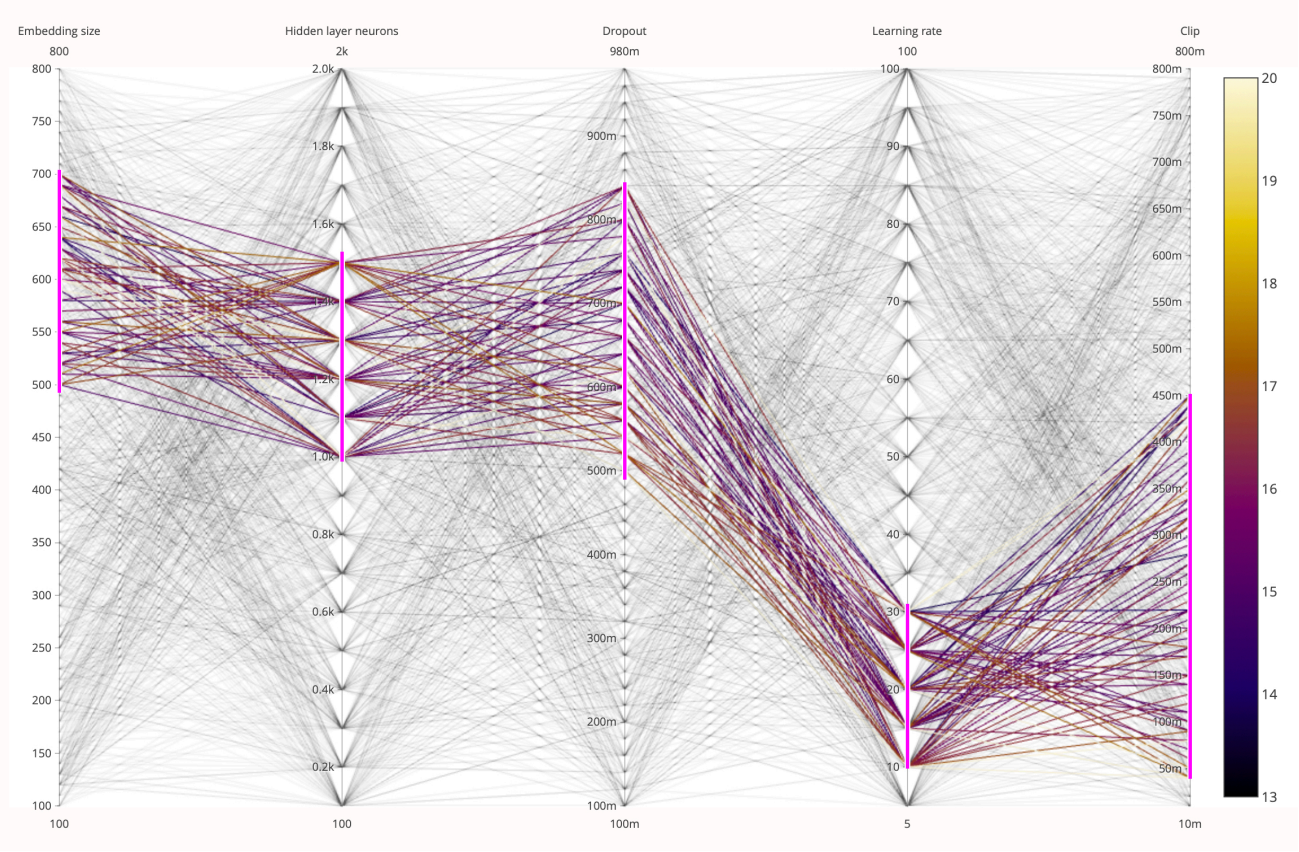


Figure 5. Hyperparameters optimization. The colorbar contains validation perplexity values.

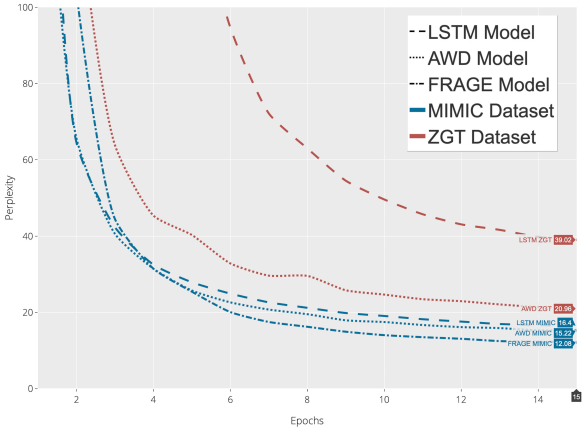


Figure 6. Perplexity on validation set

to 87% is when trying to estimate the original sentences. Sometimes, radiologists argue that original sentences are generated because they have misspellings or do not makes sense. Using the evaluation of the radiologists, we cannot state that the generated reports cannot be distinguished from original reports, although radiologists feedback was very good regarding the accuracy of the predictions.

The possible methods of accomplishing the task of cancer diagnosis using breast imaging are defined using existing

Table 5. Process methods and evaluation

Method	Accuracy A				Speed S				Doctors D				RPE
Completely human	✓	✓	✓	✓	✓				✓	✓	✓	✓	5
Completely AI	✓	✓			✓	✓	✓	✓	✓				7
Semi-automated	✓	✓	✓	✓	✓	✓			✓	✓	✓		7
Semi-automated+	✓	✓	✓	✓	✓	✓	✓		✓	✓			**9**

automation, explained in detail in Related Work (Section 2.1). For this study, we set the following possibilities:

1. Completely human. In this scenario, all the steps are done by radiologists.

2. Completely automatic. The automation is done using work of Zhu et al [26]. In this case it is a end-to-end approach, where the steps of unstructured and structured report writing are not done at all. The method takes the image as input and outputs the diagnosis.

3. Semi automatic. This case can be implemented with the existing state of the art methods in the field [20, 23]. In this approach, the writing of the unstructured report by medical stuff is the only step that restricts a full automation.

4. Semi automatic plus. This process is similar with Semi Automatic, but it's made more efficient with the addition of the real world implementation of EnDuRLM (plus).

$$RPE = 2 \cdot A + S - D \quad (2)$$

To evaluate these methods, we defined our own metric called Radiology Process Evaluation (RPE). The metric takes into consideration three aspects:

1. Accuracy (A) - evaluates the correctness of the diagnosis.
2. Speed (S) - evaluates the time of the process.
3. Doctor (D) - evaluates the number of doctors involved in the process.

All aspects have a value between 1 and 4, 1 minimum and 4 maximum for accuracy and speed, and reversed scale for doctor. The ideal case would imply that accuracy is the highest, speed is the highest and doctor involvement is the lowest. According to the equation 2, we doubled the accuracy in the calculation, because it reflects the quality of the diagnosis.

According to the results of RPE, showcased in Table 5, the addition proposed by us with EnDuRLM to the existing method increases the RPE value with 2 points, through improvements in speed and doctor sections. A high value of RPE means a better process. The maximum that can be achieved is 11 and can be achieved with one step improvements in the speed and doctor sections.

6. CONCLUSION AND DISCUSSIONS

In this paper we explore and extend the related work on word modeling process for medical applications. As a main theoretical contribution, first, we introduce a new approach for English Dutch Radiology Language Modelling, named by us EnDuRLM. EnDuRLM is equipped with the state-of-the-art machine learning models for Radiology Language Modelling, and is able to transfer the knowledge from English to Dutch language. Secondly, we devise an appropriate metric named Radiology Process Evaluation(RPE). En-DuRLM was systematically tested in order to obtain the optimal parameters. The metrics used to asses the En-DuRLM performance were perplexity and RPE. Additional evaluation has been done by radiologists.

In the context of word embeddings (Section 3.1) we made three key observations. Firstly, we discovered that word embeddings encode linguistic regularities and patterns which can be represented as linear translations of word vectors. Second, we found that the words (or more precisely the word vectors) closest to some word in the embedding space are semantically similar to that word. Third, we observed that medical relationships are encoded in the embeddings.

Future developments include more optimization, but also using the framework for other languages that have similarities. Besides word prediction, using these models, the following applications can be developed in real world: missing data recovery, quality check and educational tool.

7. ACKNOWLEDGEMENTS

First, I want to thank Christin Seifert and Elena Mocanu for supervising and enabling me to perform and write my research paper. I want to thank Jeroen Geerdink, the innovation manager from Hengelo Hospital (ZGT) for facilitating the ZGT dataset and MRON radiologists: Jeroen Veltman, Rob Bourez, Chrit Stassen, Frank Wesseling, Onno Vijlbrief for their expertise in the evaluation of generated text.

8. REFERENCES

[1] Breast imaging reporting and data system. *BI-RADS Committee, American College of Radiology*, 1998.

[2] M. A. Nogueira, P. Henriques Abreu, P. Martins, P. Machado, H. Duarte, and J. Santos. Image descriptors in radiology images: a systematic review. *Artificial Intelligence Review*, 47, 06 2016.

[3] M. Aswathy and M. Jagannath. Detection of breast cancer on digital histopathology images: Present status and future possibilities. *Informatics in Medicine Unlocked*, 8:74 – 79, 2017.

[4] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, 2018.

[5] M. Broeders, S. Moss, L. NystrÃűm, S. Njor, H. Jonsson, E. Paap, N. Massat, S. Duffy, E. Lynge, and E. Paci. The impact of mammographic screening on breast cancer mortality in europe: A review of observational studies. *Journal of Medical Screening*, 19(1_suppl):14–25, 2012. PMID: 22972807.

[6] S. Cornegruta, R. Bakewell, S. Withey, and G. Montana. Modelling radiological language with bidirectional long short-term memory networks. *CoRR*, abs/1609.08409, 2016.

[7] K. Fortney. Pre-processing in natural language machine learning, Nov 2017.

[8] C. Gong, D. He, X. Tan, T. Qin, L. Wang, and T. Liu. FRAGE: frequency-agnostic word representation. *CoRR*, abs/1809.06858, 2018.

[9] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, may 2016.

[10] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2000.

[11] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Merida, C. I. Sanchez, R. Mann, G. Heeten, and N. Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, 35, 08 2016.

[12] B. Krause, E. Kahembwe, I. Murray, and S. Renals. Dynamic evaluation of neural sequence models. *CoRR*, abs/1709.07432, 2017.

[13] R. Lacson and R. Khorasani. Natural language processing for radiology (part 2). *Journal of the American College of Radiology*, 8(8):583 – 584, 2011.

[14] C. D. Lehman, R. D. Wellman, D. S. M. Buist, K. Kerlikowske, A. Tosteson, and D. L. Miglioretti. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 175 11:1828–37, 2015.

[15] T. Magnuson and S. Hunnicutt. Measuring the effectiveness of word prediction: The advantage of long-term use. *TMH-QPSR*, 43, 01 2002.

[16] G. Melis, C. Dyer, and P. Blunsom. On the state of the art of evaluation in neural language models. *CoRR*, abs/1707.05589, 2017.

[17] S. Merity, N. S. Keskar, and R. Socher. Regularizing

and optimizing LSTM language models. *CoRR*, abs/1708.02182, 2017.

[18] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[19] S. Pathak, J. v. Rossen, O. Vijlbrief, J. Geerdink, C. Seifert, and M. van Keulen. Post-structuring radiology reports of breast cancer patients for clinical quality assurance. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

[20] S. Pathak, J. Van Rossen, O. Vijlbrief, J. Geerdink, C. Seifert, and M. Van Keulen. Automatic structuring of breast cancer radiology reports for quality assurance. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2018-November:732–739, 2019.

[21] J. M. Paulett and C. P. Langlotz. Improving language models for radiology speech recognition. *Journal of Biomedical Informatics*, 42(1):53 – 58, 2009.

[22] A. Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *CoRR*, abs/1808.03314, 2018.

[23] D. A. Sippo, G. I. Warden, K. P. Andriole, R. Lacson, I. Ikuta, R. L. Birdwell, and R. Khorasani. Automated extraction of bi-rads final assessment categories from radiology reports with natural language processing. *Journal of Digital Imaging*, 26(5):989–994, Oct 2013.

[24] R. Smith-Bindman, D. L. Miglioretti, R. Rosenberg, R. J. Reid, S. H. Taplin, B. M. Geller, and K. Kerlikowske. Physician Workload in Mammography. *American Journal of Roentgenology*, 190(2):526–532, feb 2008.

[25] Z. Yang, Z. Dai, R. Salakhutdinov, and W. W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. *CoRR*, abs/1711.03953, 2017.

[26] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention âĹŠ MICCAI 2017*, pages 603–611, Cham, 2017. Springer International Publishing.

APPENDIX

A. EXAMPLE OF DATASETS REPORTS BEFORE AND AFTER DATA CLEANING

MIMIC

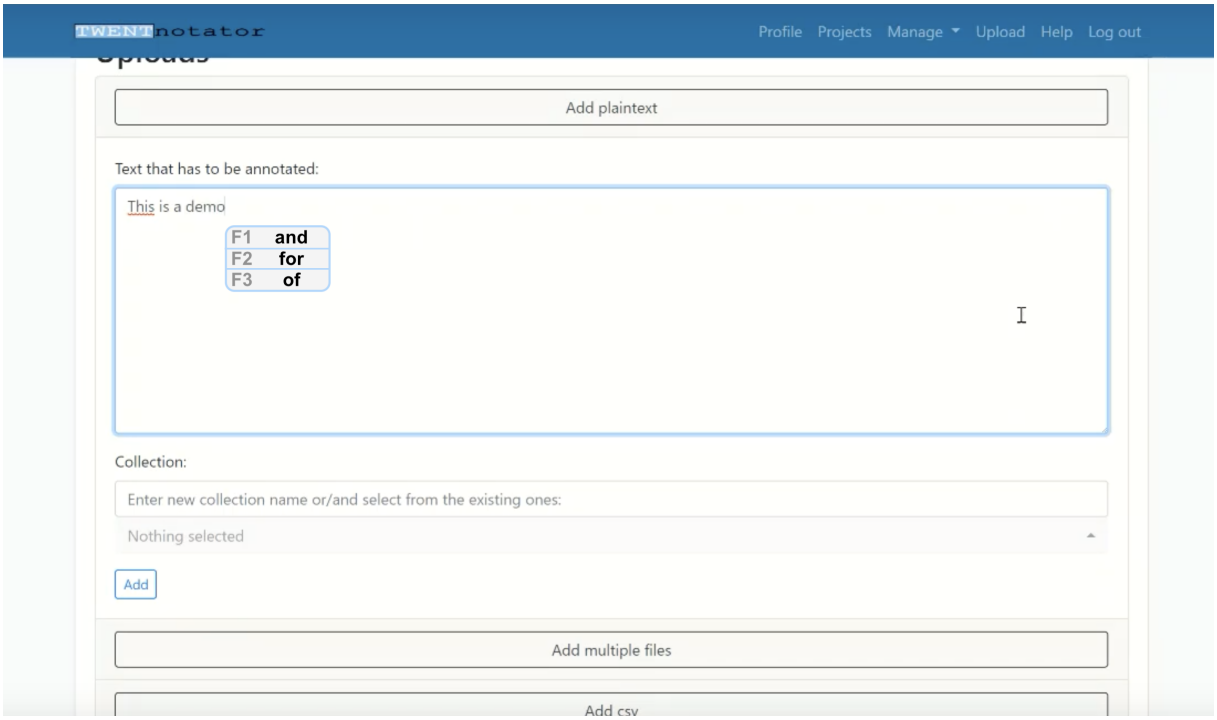
Before	After
[**2114-6-26**] 8:24 AM MAMMOGRAM (SCREENING); CAD SCREENING Clip # [**Clip Number (Radiology) 60955**] Reason: SCREENING FINAL REPORT INDICATION: Screening, remote negative left breast biopsy. Nulliparous patient. FILM-SCREEN MAMMOGRAPHY: Additional view obtained of each breast. Fatty parenchyma. Pacing device overlies right axilla. Breast tissues demonstrate diffuse scattered fibroglandular opacities without primary or secondary sign of malignancy or interval change from [**2181-12-6**]. No new suspicious masses, clusters of microcalcifications, developing areas of density or architectural distortion are seen. IMPRESSION: No evidence of malignancy. BIRADS 2 - benign findings.	screening remote negative left breast biopsy <end> nulliparous patient <end> additional view obtained of each breast <end> fatty parenchyma <end> pacing device overlies right axilla <end> breast tissues demonstrate diffuse scattered fibroglandular opacities without primary or secondary sign of malignancy or interval change from <unk> <end> no new suspicious masses clusters of microcalcifications developing areas of density or architectural distortion are seen <end> no evidence of malignancy <end> birads N benign findings <end>

ZGT

Before	After
Medische gegevens, Routine in verband met lipo vulling, Tepel uitvloed, gelig - melkachtig, Mammapathologie uitsluiten, Mammografie beiderzijds: Vergeleken wordt met 13/09/2123, Bekende asymmetrie van het retromamillaire klierweefsel met rechts meer weefsel dan links, Geen suspecte massa's of densiteiten, Geen suspecte clusters microcalcificaties, Binnenmembraan rechts laat linguini sign zien, Redelijk beoordeelbaar mammogram bij dens fibroglandulair weefsel ACR classificatie-III, Geen pathologische klieren axillair, Conclusie, Normaal mammogram, BIRADS-I,	medische gegevens <end> routine in verband met lipo vulling <end> tepel uitvloed <end> gelig melkachtig <end> mammapathologie uitsluiten <end> vergeleken wordt met <unk> <end> bekende asymmetrie van het retromamillaire klierweefsel met rechts meer weefsel dan links <end> geen suspecte massa s of densiteiten <end> redelijk beoordeelbaar mammogram bij dens fibroglandulair weefsel acr classificatie N <end> geen suspecte clusters microcalcificaties <end> binnenmembraan rechts laat linguini sign zien <end> geen pathologische klieren axillair <end> normaal mammogram birads N <end>

*the reports displayed here are composed from original reports and do not contain real data, thus no privacy issues

B. NEXT WORD PREDICTOR FEATURE - MOCK-UP



C. GENERATED TEXT USING BEST MODEL FOR EACH LANGUAGE

#	MIMIC
1	Screening analog mammography with icad computer aided detection the breasts are predominantly fatty.
2	There is a N cm opacity seen in the upper outer quadrant.
3	There are vascular calcifications however a single bb likely represents a small lymph node in the right retroareolar region.
4	There is a N cm opacity seen in the upper outer quadrant.
5	No evidence of malignancy. Birads N benign findings.
#	**ZGT**
1	Goed beoordeelbaar dens klierweefsel rechts mamma compositiebeeld b bij densiteit geen pathologische laesies of stellate laesies.
2	Ter nadele van het tijdsinterval gebied leeg denser als rondom ieder N en drainage aanvullend.
3	Echografie net boven van de areola thans een massa zichtbaar met littekenweefsel van N cm.
4	Dientengevolge bevindingen op de mlo opname subcutis lateraal craniaal in de rechtermamma scherp densiteit gezien met niet positie.
5	Birads N geen linker uit geesteren voelt sinds N maand knobbeltje in de borst nlateraal.

D. EVALUATION FORM TABLES FOR HOSPITAL

#	Sentence	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Motivation/Comments
ENGLISH: MIMIC DATA SET							
1	No pathologically enlarged lymphadenopathy identified in the left axilla.	○	○	○	○	○	
2	There are vascular calcifications however a single bb likely represents a small lymph node in the right retroareolar region.	○	○	○	○	○	
3	Screening analog mammography with icad computer aided detection the breasts are predominantly fatty.	○	○	○	○	○	
4	Breast tissue is heterogeneously dense limiting the sensitivity of mammography and there are no suspicious masses calcifications or changes since <unk>.	○	○	○	○	○	
5	The patient was informed that these calcifications are quite posterior and if they cannot be biopsied with stereotactic guidance then wire localization and surgical excision will be recommended.	○	○	○	○	○	

#	Sentence	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Motivation/Comments
DUTCH: ZGT DATASET							
1	Status na mst rechts met operatieclips volumeverlies postoperatieve fibrose distorsie en huidverdikking bij status na radiotherapie.	○	○	○	○	○	
2	Ook deze komt overeen met een langer bestaande afwijking op het mammogram.	○	○	○	○	○	
3	Goed beoordeelbaar dens klierweefsel rechts mamma compositiebeeld b bij densiteit geen pathologische laesies of stellate laesies.	○	○	○	○	○	
4	Ter nadele van het tijdsinterval gebied leeg denser als rondom ieder N en drainage aanvullend.	○	○	○	○	○	
5	Beeld van tweetal intramammaire lymfeklieren in het laterale bovenkwadrant van de linkermamma birads classificatie N.	○	○	○	○	○	