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

Mihai David Marin
*Department of Computer Science*
*University of Twente*
Enschede, the Netherlands
marinmihaidavid97@gmail.com

Elena Mocanu
*Department of Computer Science*
*University of Twente*
Enschede, the Netherlands
e.mocanu@utwente.nl

Christin Seifert
*Department of Computer Science*
*University of Twente*
Enschede, the Netherlands
c.seifert@utwente.nl

***Abstract*—Radiologists are required to write free paper text reports for breast screenings in order to assign cancer diagnoses in a later step. The current procedure requires considerable time and needs efficiency. In this paper, to streamline the writing process and keep up with the specific vocabulary, a word prediction tool using neural language models was developed. Consequently, challenges as different languages (English, Dutch), small data sizes and low computational power have been overcome by introducing a novel English-Dutch Radiology Language Modelling process. After defining model architectures, the process involves data preparation, bilevel hyperparameters optimization, configuration transfer and evaluation. The model is able to improve the current workflow and successfully meet the computational constraints, based on both an intrinsic and extrinsic evaluation. Given its flexibility, the model opens the door for future research involving other languages and also an extensive set of real-world applications.**

***Index Terms*—mammography, medical reports, neural language model, text generation, natural language processing**

## I. Introduction

Breast cancer is the most common 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 [1]. 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 [2], 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% [3].

After applying medical 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, such as information retrieval [4], reports structuring [5] or diagnosis classification [6]. Such tools improve the speed of the process, the accuracy of the diagnosis, and at the same time, reduce the number of clinicians 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 [7]. 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 have shown that word prediction increased the text composition time by at least 22% in long term use [8].

The main contributions of this paper are:

1) The process *English Dutch Radiology Language Modelling (EnDuRLM)*, developed to overcome challenges such 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 Long short-term memory (LSTM) architecture [9]. Then, the configuration is transferred to the Dutch dataset and the other model architectures: Averaged stochastic gradient weight-dropped LSTM (AWD-LSTM) [10] and Frequency-agnostic word embedding (FRAGE) [11]. In the end, the models are evaluated using perplexity, and the 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 a 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 the next word suggestion, where the model predicts the upcoming word based on the 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 in a vocabulary and structure specific manner.

This paper is structured as follows: first, we present related work. Second, we describe the datasets we have used. Then, we explain the EnDuRLM method in detail. Finally, we present the

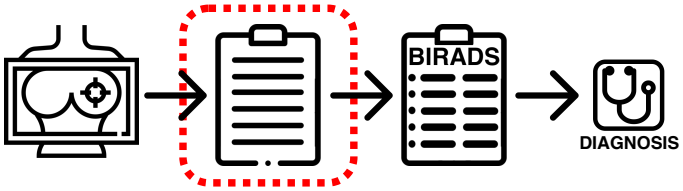


Figure 1. General flow of information for cancer diagnosis using breast imaging.

results, draw conclusions and discuss future work. The trained models, and source code are available on Github-[1].

## II. Related work

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

### A. Breast imaging automation

Because radiologists interpret 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) [12]. The BI-RADS lexicon provides specific terms to be used to describe findings, but also describes the desired report structure, for example, a report should contain breast composition and a clear description of findings [5].

Diagnosing cancer based on radiology images involves writing a report that describes the observations on the images. Those reports are usually unstructured and NLP-based post-processing can be used to obtain a structured report [13]. In the end, a clinicians determine the diagnosis based on the unstructured report, or the structured report if available. The complete process is illustrated in Figure 1.

In the last decade, Computer-Aided Diagnosis solutions [14] 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 [15], and therefore their accuracy should be improved to be ultimately considered useful. These limitations are overcome by development of deep learning solutions, which have been shown to achieve near-human performances for some applications [16]. If previous methods rely on regions of interest (parts of the image), Zhu et al. [17] proposed an end-to-end approach based on the whole mammogram. The results were modest compared to classic methods, and are still too restricted in their performance 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 [18]. 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 [6].

More recently, in 2018, using a Conditional Random Field (CRF) model, Pathak et al. [5], [13] developed an algorithm which structures free-text dutch radiology reports on breast cancer for quality assurance. The results close to clinicians' accuracy (approximately 95%) allowed clinical implementation as an annotation tool (TWENTnotator) where an unstructured report is automatically converted into a BI-RADS structured report.

### B. 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 by Sherstinsky's paper [9]. 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 [10]. 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. [11] makes use of a way to learn frequency-agnostic word embedding (FRAGE) using adversarial training. This representation technique is build at its own on top of a joint improvement of other's work: AWD-LSTM-MoS [19] and dynamic evaluation [20].

## III. Data sets

This section describes the ZGT and MIMIC-III data sets, 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. Both datasets are provided under a data use agreement and were subject to de-identification by removing privacy-sensitive patient data such as id, name, address, date of birth, etc. We use the following two datasets.

1) ZGT (Hospital Group Twente) Mammography Reports Database in the Dutch Language. This database provides approximately 48,000 reports dated from 2012 to 2017.
2) MIMIC-III (Medical Information Mart for Intensive Care) developed by MIT Lab for Computational Physiology [21] in the 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 further refer to this dataset as MIMIC.

### A. Word Embeddings Overview

According to similar studies [22], a qualitative assessment of the word embeddings obtained by training a Continuous Bag Of Words model on the MIMIC dataset has been made.

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

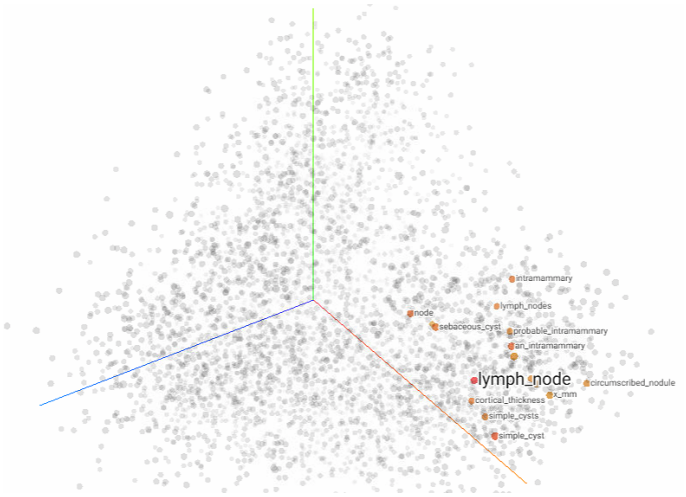


Figure 2. Visualization of the embedding space for MIMIC corpus learnt by a Word2Vec model, highlighting nearest neighbours of the word *lymph*.

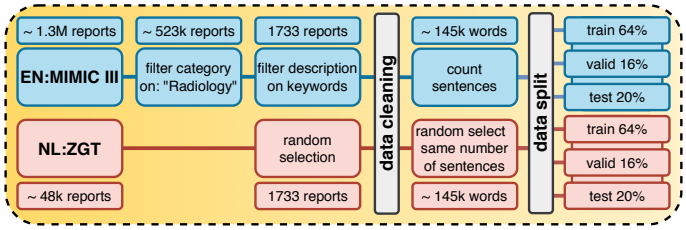


Figure 3. Flow of data in preprocessing step of EnDuRLM

The 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 embeddings can be found in Figure 2. The embeddings were visualized by means of Principal Component Analysis (PCA) [23], using the top three principal components to reduce the dimensionality of the dataset to three dimensions. As described by Mikolov et al. [24] 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", and "fluid collection".

### B. Data preparation

Preprocessing the text is a crucial task requiring optimal tools given both the data and language models [25]. A typical approach for data preparation is as follows: each dataset first has to be manually reviewed in order to apply cleaning of irrelevant data as, such as headings, punctuation, strange names and quoted dialogue sequences. The next step is lower casing all the words. Tokenization is the process of dividing text into words and sentences. Figure 3 illustrates the process step by step.

*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

Table I
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 |

structure and morphology, we decided to have the same number of sentences in both datasets.

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, 52,3000 reports. The description of the reports varied from a 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 I. Filtering again based on these descriptions resulted in almost two 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, and therefore approximately the same number of words: almost 150 thousand.

*2) Data cleaning:* Given the differences in layout and structure, the data cleaning was performed separately for each dataset, 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.
- Personal data in MIMIC is replaced by tag $\langle unk \rangle$ (for unknown).
- ZGT reports use commas as sentence delimiters. Those commas were replaced with full stops.
- In both datasets, text was split into sentences, such that each row corresponds to one sentence.
- Measurements values, date and time were also replaced by $\langle unk \rangle$.
- Common typos were fixed and common abbreviations were expanded (for e.g. y.o. to years old or dr. to doctor).
- Auto generated headers were deleted.
- Remove all punctuation, double spaces and tabs.
- Split in sentences using Spacy.
- All sentences were converted to lower case.
- Numbers and roman numerals were replaced by $N$.
- End of each line was replaced by $\langle eos \rangle$.

We keep stop words, because of their relevance in the final model. Table II showcases an example of each dataset before and after data cleaning.

*3) Data preprocessing:* The file containing the same number of sentences of cleaned text (retrieved randomly) was split into 80% train and 20% test. The training set was further split into 80% train and 20% validation. These decisions were taken in accordance with [26]. Taking into account the small

Table II
EXAMPLE OF REPORTS BEFORE AND AFTER DATA CLEANING, FOR BOTH, ZGT AND MIMIC DATASET

| MIMIC | |
|---|---|
| **Before** | **After** |
| [**2114-6-26**] 8:24 AM<br>MAMMOGRAM (SCREENING); CAD SCREENING Clip # [**Clip Number (Radiology) 60955**]<br>Reason: SCREENING<br>FINAL REPORT<br>INDICATION: Screening, remote negative left breast biopsy. Nulliparous patient.<br><br>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.<br><br>IMPRESSION: No evidence of malignancy. BIRADS 2 - benign findings. | screening remote negative left breast biopsy <end> nulliparous patient <end><br><br>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><br><br>no evidence of malignancy <end> birads N benign findings <end> |
| **ZGT** | |
| **Before** | **After** |
| Medische gegevens,<br>Routine in verband met lipo vulling,<br>Tepel uitvloed, gelig - melkachtig, Mammapathologie uitsluiten,<br><br>Mammografie beiderzijds:<br>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,<br><br>Conclusie,<br>Normaal mammogram, BIRADS-I, | medische gegevens <end><br>routine in verband met lipo vulling <end><br>tepel uitvloed <end> gelig melkachtig <end> mammapathologie uitsluiten <end><br><br>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><br>geen pathologische klieren axillair <end><br><br>normaal mammogram birads N <end> |

*the reports displayed here are artificially constructed from original reports, but close to original reports and do not contain real data.

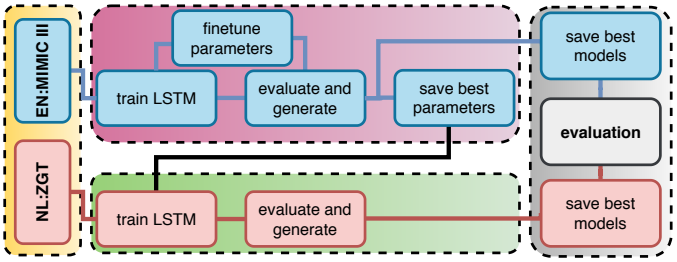


Figure 4. Flow diagram of the proposed English-Dutch Radiology Language Modelling (EnDuRLM) process

ZGT dataset size, we decided to include all the words in the vocabulary. The sizes of the vocabularies were: 3148 for MIMIC and 4443 for ZGT.

## IV. ENGLISH-DUTCH RADIOLOGY LANGUAGE MODELLING

This section describes our systematic approach to solve the given problem, and its real-world implementation scenario.

The English-Dutch Radiology Language Modelling (EnDuRLM) framework is designed to overcome challenges, such 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, the hyperparameters optimization is done on the LSTM architecture and the MIMIC data set using a model-free bilevel optimization search. The best configuration is transferred to the other models (AWD-LSTM and FRAGE) and languages (Dutch: ZGT). Finally, we evaluate and compare all models. For a better overview, the EnDuRLM process is illustrated in Figure 4, further details follow in the next sections.

*a) EnDuRLM - Model architecture:* The LSTM architecture is designed to be better at storing information and finding and learning long-term dependencies than standard recurrent networks. Research from the last years has proved that well-tuning LSTM baseline model outperform high-level architectures in the field of word-level language modelling [27].

We use three model architectures: LSTM [9], AWD-LSTM [10] and FRAGE-LSTM [11]. The code for these implementations is retrieved from open source repositories and adjusted to fit EnDuRLM code. According to previous studies [26], we used a batch size of 20 and unrolled the network for 35 time steps.

*b) EnDuRLM - Bilevel hyperparameters optimization:* All networks were trained with Stochastic Gradient Descent (SGD). Merity et al. [10] 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 parameters shown in Table III in the ranges specified in Merity et al. [10]. The second round of random search consisted of restricting the ranges based on the top 10% models regarding values of perplexity from the first random search. The best configuration is stored and used for the other model architectures in the configuration transfer phase.

*c) Configuration transfer to models and language:* We applied the best hyperparameters setting on the remaining models for the MIMIC dataset and all models for ZGT dataset. Taking into account that AWD-LSTM and FRAGE models contain additional hyperparameters for the added features, for this study we will keep their default values (dropout for RNN layers is 0.3; for input embedding layers is 0.65; to remove

Table III
HYPERPARAMETER 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 | 50.00% |
| Total | | | | **67.20%** |

*the step size is the same for both searches

words from embedding layer is 0.1; and to the RNN hidden-to-hidden matrix is 0.5. Furthermore, alpha L2 regularization on RNN activation is 2; beta slowness regularization applied on RNN activation is 1; weight decay is $1.2 \cdot e^{-6}$) for a similar dataset, as described in [10], [11].

*d) Intrinsic Evaluation:* Intrinsic evaluation metrics allow us to measure the quality of a model independent of a particular application [28]. The most typical metric used to measure the efficiency of a language model is perplexity.

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

where: $w_1, w_2, ..., w_N$ are the words from the test set $W$ with length $N$. 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. Consequently, the worst model would have a perplexity equal to the size of vocabulary, 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, resulting in a better model.

*e) Extrinsic evaluation:* 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. In this context, the extrinsic evaluation refers to the integration of our proposed language model in the TWENTnotator application and measuring how much the application improves [28]. Implementation of the language models as word prediction in TWENTnotator will facilitate extrinsic evaluation. Moreover, samples of the generated text are sent to radiologists for evaluating correctness and relevance of the generated texts.

## V. EXPERIMENTS AND RESULTS

*a) Bilevel hyperparameters optimization:* Because of computational resources limitations for the ZGT (no graphical card computation power on the server where the data had to be processed for security reasons), we optimized hyperparameters using a bilevel random search, i.e., two stages of random search where the set of hyperparameters for the second stage is determined by the outcome of the first stage, on MIMIC dataset.

In the first round of random search, we trained and tested 1000 different LSTM models with the parameter ranges defined in Table III. Figure 5 illustrates model parameters, highlighting the 100 best performing 100 models w.r.t. perplexity on the test set. This highlighting is used to constrain the hyperparameter ranges for the second round of random search. The reason for a second search is to ensure robustness and increase accuracy. We use the same the step size, while ranges were reduced by an average of 67.2% according to Table III. This round of random search trained and tested 100 different LSTM models and the final best performing model w.r.t.validation set perplexity, had the following configuration: 1400 hidden neurons; an embedding size of 660; learning rate of 30 and a dropout of 0.76, and a gradient clipping norm of 0.28.

Table IV
PERPLEXITY VALUES FOR VALIDATION AND TEST SET ON BOTH DATASETS

| Dataset | Model | Epochs [#] | Validation perplexity | Test perplexity |
|---|---|---|---|---|
| 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** |

*b) Transfer learning:* The best performing configuration was applied to the other model architectures (AWD-LSTM and FRAGE) for both datasets. The models are again evaluated using the perplexity metric, and the results are shown in Table IV. Comparing the results for the MIMIC dataset, we observe an improvement of perplexity in validation set of almost 35% from LSTM to FRAGE, which is quite promising. 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-LSTM, and we cannot provide an exact result with FRAGE. The problems encountered when trying to implement FRAGE were technical[3]. We can estimate the value of FRAGE for ZGT by assuming the same improvment of 12% improvement from AWD-LSTM to FRAGE on MIMIC. The differences between English and Dutch are reasonable, 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 architecture (LSTM, AWD-LSTM, FRAGE) and each data set (MIMIC, ZGT), resulting in five 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

[2]https://github.com/yannislinardos/annotationTool, accessed Sep, 2020

[3]We had no access to graphical power unit capable of running CUDA environment

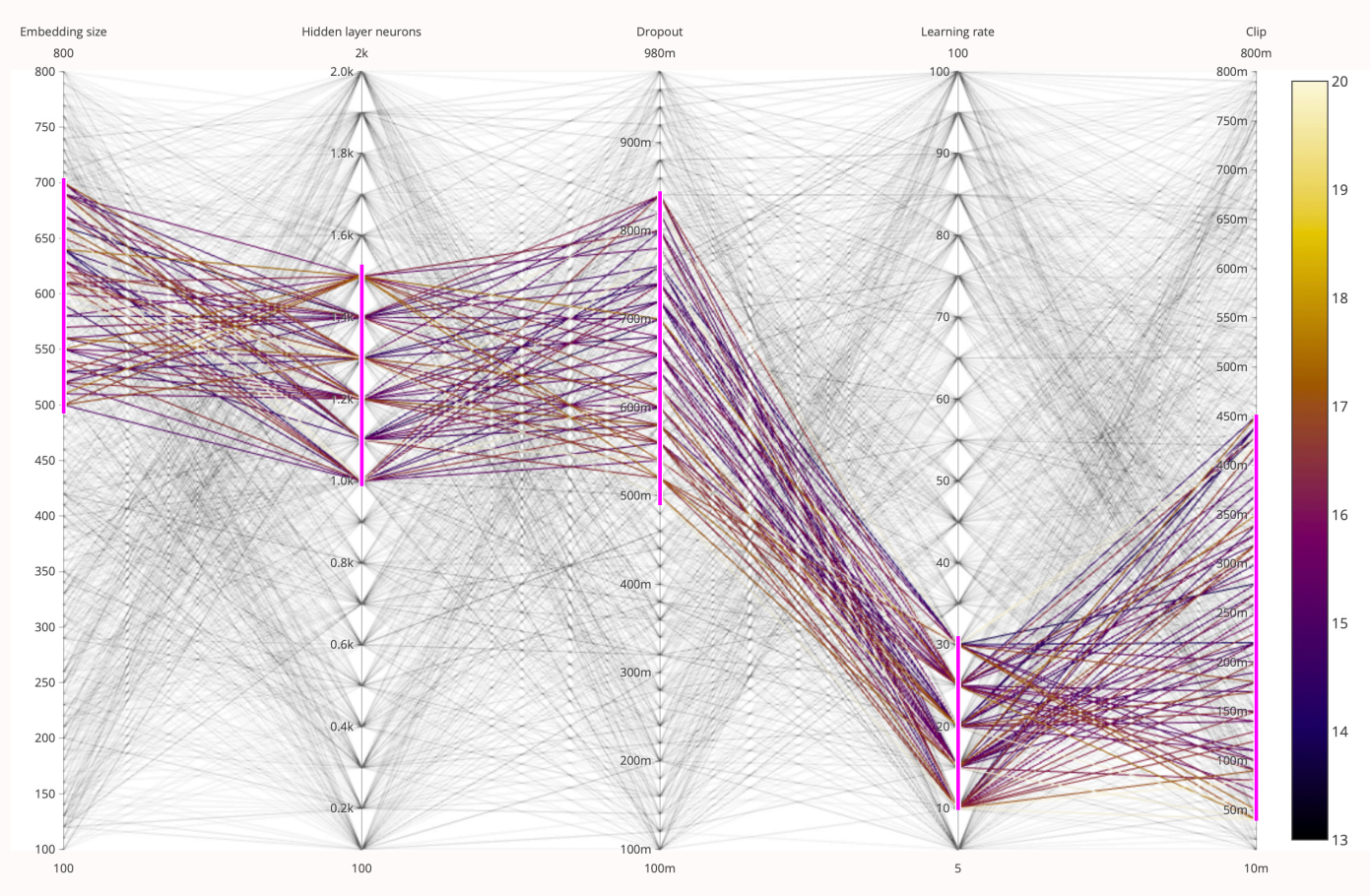


Figure 5. Parallel plot of hyperparameter optimization results. Color indicates validation perplexity values. 1000 models are shown in total, the best 100 are shown in color.

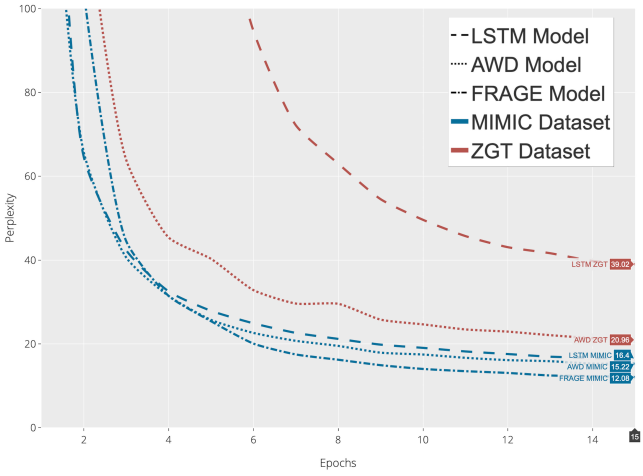


Figure 6. Perplexity on validation set for both datsets, MIMIC (blue) and ZGT (red), using LSTM, AWD-LSTM, and FRAGE models.

100. Figure 6 shows high convergence for each setup. The perplexity decreases under 100 in the first three epochs and stabilizes quite fast for each setup, excluding the LSTM model with ZGT, where the perplexity shows a slower convergence.

We analyzed five sentences generated with the best setup as depicted in Table VII, and the results seem sensible and the order of the words is very good. The generated texts are further analyzed by domain experts to assess quality of the generated reports.

*c) EnDuRLM: Extrinsic evaluation:* Using the best architecture of each language (FRAGE for English, AWD-LSTM for Dutch), we created an evaluation form for ZGT hospital. The process involves the expertise of the radiologists, because of their extensive knowledge about specific vocabulary and the domain. In the evaluation, we show either sentences automatically generated sentences or sentences retrieved from an original report. We asked clinicians to estimated the source of the report on a Likert scale. Strongly Disagree means that the sentences was retrieved from original report, Neutral means the source of the sentence is unsure, and Strongly Agree means the text was perceived as being generated automatically. The sentences are randomly chosen without any connection between them, and are shown in Table VII. We also asked for qualitative feedback. With this evaluation technique, we wanted to investigate whether clinicians can distinguish between a and generated reports and find the reasons for their decisions.

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

Table V
RADIOLOGISTS EVALUATION, FOR A TOTAL NUMBER OF THREE ORIGINAL SENTENCES AND TWO GENERATED SENTENCES.

| Radiologist | Original | | Generated | | Overall accuracy [%] |
|---|---|---|---|---|---|
| | MIMIC | ZGT | MIMIC | ZGT | |
| 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 |

overview, we decided to count the number of correct guesses and illustrate the results in Table V. All radiologists correctly classify the generated sentences for both languages with 100% accuracy, arguing that the sentences make a lot of sense, but some words suggest different context. The point where the overall accuracy went 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 make 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 automation, as explained in Section II. Four possibilities exist:

1) Completely human. In this scenario, all the steps are done by radiologists.
2) Completely automatic. The automation is done using the work of Zhu et al. [17]. In this case, it is an end-to-end approach, and does not involve writing any report. 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 [6], [13]. The fact that radiologists unstructured texts in practice is the only step that restricts a full automation.
4) Semi-automatic plus. This process is similar to Semi-Automatic, but it is made more efficient with the addition of the real-world implementation of EnDuRLM.

To evaluate these methods, we defined our metric called Radiology Process Evaluation (RPE). The metric takes into consideration three aspects: i) Accuracy (A) - evaluates the correctness of the diagnosis; ii) Speed (S) - evaluates the time of the process, and iii) Doctor (D) - evaluates the number of doctors involved in the process.

$$RPE = 2 \times A + S - D \tag{2}$$

All aspects have a value between 1 and 4, 1 minimum and 4 maximum for accuracy and speed, and reversed scale for doctor. In the ideal case, accuracy has the highest value (4), speed has the highest value (4) and doctor involvement has the lowest value (1). Values 2 and 3 represent values close to minimum and maximum. We propose the following categorisation for the scales: 1 - very bad (0-25% for accuracy), very slow (more than 1 hour for speed), minimum (zero doctors for number of doctors involved), 2 - bad (25-50% for accuracy), slow (between 30 and 60 minutes for speed), less (1-2 doctors for number of doctors involved), 3 - good (50-75% for accuracy), fast (between 5 and 30 minutes for speed), lot (3-5 doctors for number of doctors involved), 4 - very good (75-100% for accuracy), very fast (less than 5 minutes for speed), maximum (more than 5 doctors for number of doctors involved). According to equation 2, we double the impact of accuracy in the calculation, because it reflects the quality of the diagnosis. According to the results of RPE, showcased in Table VI, the addition proposed by us with EnDuRLM increases the RPE value with 2 points, through improvements in speed and doctor metrics. A high value of RPE means a better process. The maximum that can be achieved is 11 and can be accomplished with one step improvements in the speed and doctor metrics.

Table VI
SUMMARY OF EXISTING PROCESS METHODS AND EVALUATION OF RADIOLOGY REPORTS. OUR PROPOSED METHOD IS HIGHLIGHTED IN GREEN.

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

Table VII
EXAMPLE OF GENERATED TEXT USING THE MOST ACCURATE MODEL FOR EACH LANGUAGE

| # | MIMIC |
|---|---|
| 1 | Screening analog mammography with icad computer aided detection the breasts are predominantly fatty. |
| 2 | There is a small obscurity seen in the lower inner half. |
| 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. |

## VI. CONCLUSION AND DISCUSSIONS

In this paper, we explore and extend the related work on language modeling 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, incorporates a bilevel model-free hyperparameters optimization using random search, and has the ability to perform a knowledge transfer from one architecture to another and from English to Dutch. Secondly, we devise an appropriate metric

named Radiology Process Evaluation(RPE). EnDuRLM was systematically tested in order to obtain the optimal parameters. The metrics used to asses the EnDuRLM performance was perplexity and RPE.

In the context of word embeddings, 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 may focus on scalable methods able to accommodate recently developed optimization techniques (e.g. Self-Tuning Networks [29]), but also using the framework in a more adaptive multi-task transfer learning context for other languages that have similarities. Besides word prediction, our proposed model –EnDuRLM, can also be used as a starting point for new applications in real-world, such as missing data recovery, quality check and educational tool.

## VII. Acknowledgement

We would like to thank J. Geerdink from Hengelo Hospital (ZGT) for facilitating the ZGT dataset and MRON radiologists: J. Veltman, R. Bourez, C. Stassen, F. Wesseling, and O. Vijlbrief for their evaluation of the generated texts.

## References

[1] 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*, vol. 68, no. 6, pp. 394–424, 2018.

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

[3] M. Broeders, S. Moss, L. Nyström, 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*, vol. 19, no. 1_suppl, pp. 14–25, 2012, pMID: 22972807.

[4] R. Lacson and R. Khorasani, "Natural language processing for radiology (part 2)," *Journal of the American College of Radiology*, vol. 8, no. 8, pp. 583 – 584, 2011.

[5] 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*, vol. 2018-November, pp. 732–739, 2019.

[6] 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*, vol. 26, no. 5, pp. 989–994, Oct 2013.

[7] 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*, vol. 190, no. 2, pp. 526–532, feb 2008.

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

[9] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[10] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," in *International Conference on Learning Representations*, 2018.

[11] C. Gong, D. He, X. Tan, T. Qin, L. Wang, and T.-Y. Liu, "FRAGE: frequency-agnostic word representation," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 1334–1345.

[12] B. Committee, "Breast imaging reporting and data system," *American College of Radiology*, 1998.

[13] S. Pathak, J. van 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.

[14] 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*, vol. 47, 06 2016.

[15] 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*, vol. 175 11, pp. 1828–37, 2015.

[16] 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*, vol. 35, 08 2016.

[17] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Cham: Springer International Publishing, 2017, pp. 603–611.

[18] J. M. Paulett and C. P. Langlotz, "Improving language models for radiology speech recognition," *Journal of Biomedical Informatics*, vol. 42, no. 1, pp. 53 – 58, 2009.

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

[20] B. Krause, E. Kahembwe, I. Murray, and S. Renals, "Dynamic evaluation of neural sequence models," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 10–15 Jul 2018, pp. 2766–2775.

[21] 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*, vol. 3, p. 160035, may 2016.

[22] S. Cornegruta, R. Bakewell, S. Withey, and G. Montana, "Modelling radiological language with bidirectional long short-term memory networks," in *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics, Nov. 2016, pp. 17–27.

[23] K. Pearson, "LIII. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[24] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *International Conference on Learning Representations (Workshop Poster)*, 2013.

[25] K. Fortney, "Pre-processing in natural language machine learning," November 2017.

[26] G. Melis, C. Dyer, and P. Blunsom, "On the state of the art of evaluation in neural language models," in *International Conference on Learning Representations*, 2018.

[27] S. Merity, N. S. Keskar, and R. Socher, "An analysis of neural language modeling at multiple scales," *CoRR*, vol. abs/1803.08240, 2018. [Online]. Available: http://arxiv.org/abs/1803.08240

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

[29] M. Mackay, P. Vicol, J. Lorraine, D. Duvenaud, and R. Grosse, "Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions," in *International Conference on Learning Representations*, 2019.