Journal of Digital Imaging (2020) 33:1002–1008
https://doi.org/10.1007/s10278-020-00327-z

ORIGINAL PAPER

Natural Language Processing in Dutch Free Text Radiology Reports: *Challenges in a Small Language Area Staging Pulmonary Oncology*

J. Martijn Nobel[1,2] · Sander Puts[3] · Frans C. H. Bakers[1] · Simon G. F. Robben[1,2] · André L. A. J. Dekker[3]

Published online: 19 February 2020

Abstract

Reports are the standard way of communication between the radiologist and the referring clinician. Efforts are made to improve this communication by, for instance, introducing standardization and structured reporting. Natural Language Processing (NLP) is another promising tool which can improve and enhance the radiological report by processing free text. NLP as such adds structure to the report and exposes the information, which in turn can be used for further analysis. This paper describes pre-processing and processing steps and highlights important challenges to overcome in order to successfully implement a free text mining algorithm using NLP tools and machine learning in a small language area, like Dutch. A rule-based algorithm was constructed to classify T-stage of pulmonary oncology from the original free text radiological report, based on the items tumor size, presence and involvement according to the 8th TNM classification system. PyContextNLP, spaCy and regular expressions were used as tools to extract the correct information and process the free text. Overall accuracy of the algorithm for evaluating T-stage was 0,83 in the training set and 0,87 in the validation set, which shows that the approach in this pilot study is promising. Future research with larger datasets and external validation is needed to be able to introduce more machine learning approaches and perhaps to reduce required input efforts of domain-specific knowledge. However, a hybrid NLP approach will probably achieve the best results.

Keywords Radiology · Reporting · Natural language processing · Free text · Classification system · Machine learning

Background

One of the most challenging tasks in healthcare informatics nowadays is how to improve accessibility to medical information. Especially in radiology, in which a large amount of imaging and textual data is captured. Combining all kinds of medical information can improve current medical data flow and can ensure better healthcare [1]. A good example of a complex process of combining data is tumor staging, for instance, in pulmonary oncology. A specific rule-based tumor classification system is used for proper staging of pulmonary oncology, as stated in the 8th TNM Classification of Malignant Tumors (TNM) [2, 3].

In radiology, the report is still considered the golden standard in communicating findings and is, despite several structuring efforts [4], usually still stored as free text. One of the challenges in radiology is how to (re-)use free text unstructured data of the radiological report for data mining purposes in, for instance, pulmonary tumor staging.

J. Martijn Nobel and Sander Puts contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10278-020-00327-z) contains supplementary material, which is available to authorized users.

✉ J. Martijn Nobel
martijn.nobel@mumc.nl

[1] Department of Radiology and Nuclear Medicine, Maastricht University Medical Center+, Postbox 5800, 6202 Maastricht, AZ, Netherlands

[2] School of Health Professions Education, Maastricht University, Maastricht, Netherlands

[3] Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, Netherlands

Natural Language Processing (NLP) is a promising method for extracting information from free text, and has been used in several studies to extract data from radiological reports [5]. However, most use English as a language and specific medical NLP software, such as medical extraction systems (e.g., cTAKES) [6], are not available in Dutch [5, 7].

In English, a rule-based pulmonary oncology TNM classification algorithm has already been built and trained on pathology reports with 72% accuracy on T-stage [8]. In addition, several Breast Imaging-Reporting and Data System (BI-RADS) classification approaches have been evaluated in English; the best results were obtained by using partial decision trees (PART) [9].

In Dutch, one study was published on free text mining in radiological reports using support vector machines (SVM) and conditional random fields (CRF) to structure free text data with a BI-RADS classification algorithm proposed as future work [10]. However, to our knowledge, no tumor-classification task based on radiology reports has been published in Dutch before.

This article describes a pilot study which shows the challenges to expect when extracting data from free text radiology reports in a small language area, like Dutch, in the classification of the T-stage of TNM pulmonary oncology.

Methods

Corpus Description

After ethical approval at the participating medical center, a training set was created which consisted of 47 radiological reports with pulmonary oncology that underwent a diagnostic staging procedure. The radiological reports have been constructed by several different radiologists, other than the authors, using a speech recognition tool (G2 Speech). Findings were stored as free text reports in a Radiological Information System (RIS, Agfa Healthcare). Every included report consisted of several structured sections with the following headings: clinical details, report, described modality, body part, and conclusion. This training set was used to identify the reporting content and to find appropriate synonyms, which were incorporated in the algorithm. Consecutively, a second set of 100 cases was used to validate the outcomes. Cases were included if a primary pulmonary malignancy was diagnosed using a computed tomography (CT) and the radiological report was present. Cases with two primary tumors and follow-up cases were excluded. After inclusion, T-stage was independently classified and labeled from the report by two authors (JMN and SP) according to the 8th TNM classification [2], because final T-stage was not explicitly mentioned in the report and could only be derived from findings described in the free text. The authors agreed on annotation guidelines for proper labeling. In case of discrepancy, consensus was reached between the two authors.

Algorithm Structure

Because of the limited training data available, a rule-based NLP algorithm with machine learning pre-processing steps was used in this study. In addition, we aimed to set a baseline for future work using more advanced machine or deep learning techniques. The used approach is subdivided into a pre-processing step and a processing step. The pre-processing is necessary to make the data suitable for analysis. The processing step is the actual algorithm (see Fig. 1: T-stage classifier and Table 1: Detailed example of the classification process).

Pre-Processing

A sectionizer was developed to only select relevant parts of the report. In this study, text was only searched when preceded by the headings *thorax* and *conclusion*. A consecutive cleaning step was introduced to remove speech recognition artifacts and to replace selected abbreviations by its full form. Open-source NLP software library SpaCy [11] was selected to perform sentence segmentation and number extraction using part-of-speech tagging (POS), as it includes pre-trained models for multiple languages and has been successfully applied on medical extraction tasks before [12].

Processing

By analyzing the 8th TNM classification [2], the T-stage classification was divided into three different items: *size*, *presence*, and *involvement* (see Fig. 1: T-stage classifier). All three items required extraction of relevant concepts (e.g., tumor; see Fig. 1*:* T-stage classifier). For every concept a set of synonyms and their conjugations was created (e.g., tumor; mass, lesion, etc.) to ensure a high recall in extracting concepts from included reports. The synonym sets were created by radiological domain experts using the training set, Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) [13] and their expertise.

Accordingly, the synonym sets were converted into a regular expression per concept. Depending on the item to extract (size, presence or involvement), the concepts were further processed by the algorithm in different ways.

To cover the item *size*, a measurement extractor was developed using POS recognition of NLP-library spaCy to extract tumor size. Tumor size was selected out of all numbers, when all of the following preconditions were fulfilled: the largest number, the number is part of an area expression, the number contains a unit (cm or mm), the number is not a distance measurement, and is not preceded by the concept "lymph node" (instead of "tumor").

The concepts to extract for the item *presence* were context validated; for every extracted concept context information (for instance, negations, uncertainty, and historical events) was

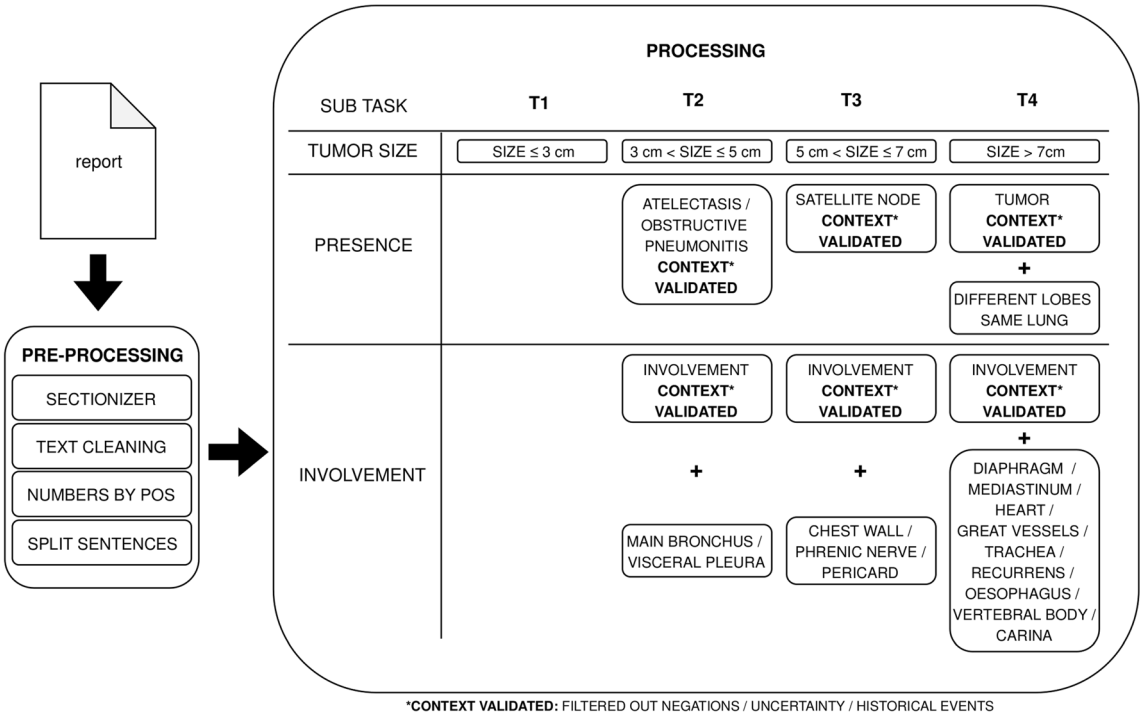

Fig. 1 T-stage classifier Schematic overview of T-stage classification. In the pre-processing step the raw data of the report is prepared for the actual processing. In the processing step tumor size extraction and a T-stage presence check of abnormalities and its involvement is performed

extracted. Only those concepts being certain by its related context were valid and used for classification.

pyContextNLP was used to extract the context including negations (modifier) related to the concept (target), as it has been translated and applied to several languages, including Dutch [14–16]. pyContextNLP has been translated and functionality has been extended to run it as a service to simplify integration with other NLP services, increasing performance and usability [17].

Finally, to extract the item *involvement*, two different concepts had to be present in same sentence: the concept "involvement" itself and, the concept being involved (e.g., possible involvement in mediastinum). The concept "involvement" is context validated; context information (for instance, negations, uncertainty, historical events) was extracted.

In addition, a specific T4-stage logic has been implemented to validate whether a tumor is present in different lobes of the same lung. Final T-stage was assigned to the most severe tumor classification found by the algorithm. A detailed example of the classification process is shown in Table 1: Detailed example of the classification process.

Results

The accuracy of the T-stage classifier on the test set was 83% ($N = 47$), and on the validation set 87% ($N = 100$) (see Table 2: T-stage classifier accuracy). Fig. 2 shows the confusion matrices of respectively the training set and the validation set, where each "*actual T-stage*" is compared with the "*predicted T-stage*". The precision (i.e., specificity), recall (i.e., sensitivity), and F_1 measure (i.e., combined metric for precision and recall) for all independent stages are obtained as shown in Table 3: Precision, recall and F_1-scores. In addition, all errors in the training set and validation set were analyzed and grouped into five specific categories with one or more subgroups: context, concepts, standardization, complexity, and spaCy (see Table 4: T-stage errors by category). In total seven errors were found in the training set and 13 in the validation set. Finally, in Appendix 1 (Concept synonyms) SNOMED concepts have been added to the table of used regular expressions, to indicate the amount of translations and synonyms missing. In Appendix 2 (Mentions related to context) and Appendix 3 (Mentions related to involvement) challenges related to context and involvement are highlighted to point out difficulties of the process.

Discussion

The aim of this paper is to gain insight in the challenges of using NLP in free text radiological reports in a small language area such as Dutch. This was done by creating an algorithm for T-stage pulmonary oncology according to the 8th TNM classification. This feasibility study is a baseline for future

Table 1 Detailed example of the classification process

RAW REPORT	PROCESSED REPORT	CLASSIFIED REPORT
Clinical details: Pulmonary malignancy? Report: CT thorax and abdomen, arterial phase Thorax: Mass visible in the left upper lobe with a maximum size estimated at image 46 of 4, 7 x 3,0 cm. Possible involvement in mediastinum. Satellite nodes visible at 8-41 with an estimated size of 1.3 cm. Lymph node visible at station 7 with a size of circa 5,2 cm. No lymph nodes visible at contralateral side. Small consolidation middle lobe. No indication of atelectasis. Abdomen: Multiple sharply edged hypodens liver lesions visible which would initially match with cysts (HU 5). Musculoskeletal No relevant findings. No metastasis. Conclusion: Tumor with satellite nodes left upper lobe	Clinical details: Pulmonary malignancy? Report: CT thorax and abdomen, arterial phase Thorax: Mass visible in the left upper lobe with a maximum size estimated at image 46 of 4, 7 x 3,0 cm. Possible involvement in mediastinum. Satellite nodes visible at 8-41 with an estimated size of 1.3 cm. Lymph node visible at station 7 with a size of circa 5,2 cm. No lymph nodes visible at contralateral side. Small consolidation middle lobe. No indication of atelectasis. Abdomen: Multiple sharply edged hypodens liver lesions visible which would initially match with cysts (HU 5). Musculoskeletal No relevant findings. No metastasis. Conclusion: Tumor with satellite nodes left upper lobe	Tumor size: **T1** (4,7 cm) Presence **T3** (satellite nodes) Involvement - Classification **T3**

DESCRIPTION

Sectionizer: filtered out sections “Thorax” and “Conclusion”

Cleaning: Colons and whitespaces within numbers removed, selected abbreviations are replaced

Size: 4,7 cm is extracted as tumor size, the number is part of an area expression, has unit cm and is not preceded by lymph node.

Presence: pyContextNLP extracted concepts and context. "Mass" and "satellite node" is found without context.

Involvement: pyContextNLP extracted "involvement" with context of type uncertainty, therefore involvement in mediastinum is ignored.

Pre-processing is performed on the raw text of the report. In the processed report, only the relevant sections remain. Every sentence in the processed report is annotated with extracted measurements, concepts (presence/involvement) and context. The final classification is obtained by the highest T-stage detected

Table 2 T-stage classifier accuracy

	Training set (N = 47)	Validation set (N = 100)
Accuracy T-stage	0.83	0.87

Accuracy scores of the training set and the validation sets

Table 3 Precision, recall and F_1-scores

Training	Precision	Recall	F_1 score
T1	0,64	1,00	0,78
T2	0,93	0,76	0,84
T3	0,70	0,78	0,74
T4	1,00	0,86	0,92
Validation	Precision	Recall	F_1 score
T1	0,90	0,82	0,86
T2	0,89	0,86	0,87
T3	0,83	0,95	0,88
T4	0,87	0,86	0,87

Precision, recall and F_1-scores for the training set and the validation set

work based on more (hybrid) advanced machine or deep learning techniques.

The described method analyzes and tries to thoroughly understand the meaning and interactions of words and phrases in the radiological report before classifying it. The main difference with a general machine or deep learning approach is that different steps are used before the final analysis is performed, instead of analyzing the report as a whole. Because the TNM classification is already rule-based, it is not necessary to force the neural network to recompose the already known T-stage rules for proper T-staging. Focusing on how to properly analyze free text was therefore one of the main goals of this approach as this can show us where difficulties can be expected and where machine or deep learning can help us smoothen this process.

The measured accuracy of this pilot study suggests that T-stage can be extracted from free text reports with a fairly high reliability. This is consistent with the earlier performed study on pathology reports written in English [8]. In addition, the strategy used for extracting the items *size*, *presence*, and *involvement* according to the 8th TNM classification seems promising. The obtained results (precision, recall, and F_1 score) for the training and validation set are in most cases at least comparable.

When looking at the pre-processing and processing steps, several important findings should be addressed. First of all, identification of synonyms of the chosen items is of utmost importance, because vocabulary used for describing tumors differs widely among reporters. This variability in vocabulary makes it difficult to use machine learning for finding appropriate synonyms at this stage, because a large amount of data is needed. However, when a sufficient amount of data is available word embeddings could be created, which might be used to automatically find synonyms for used concepts. This study highlights the importance of using domain specific knowledge when building a (rule-based) algorithm when training data is limited.

Attempts to find proper synonyms by using (the Dutch) SNOMED-CT failed. Used synonyms are not always a synonym of the proper SNOMED-CT concept, but for example, a synonym of a related super concept. Iterating over all supertype (parent) concepts is tedious and most are irrelevant (e.g., several tumor synonyms can be found searching for abnormal morphology). In addition, the Radiological

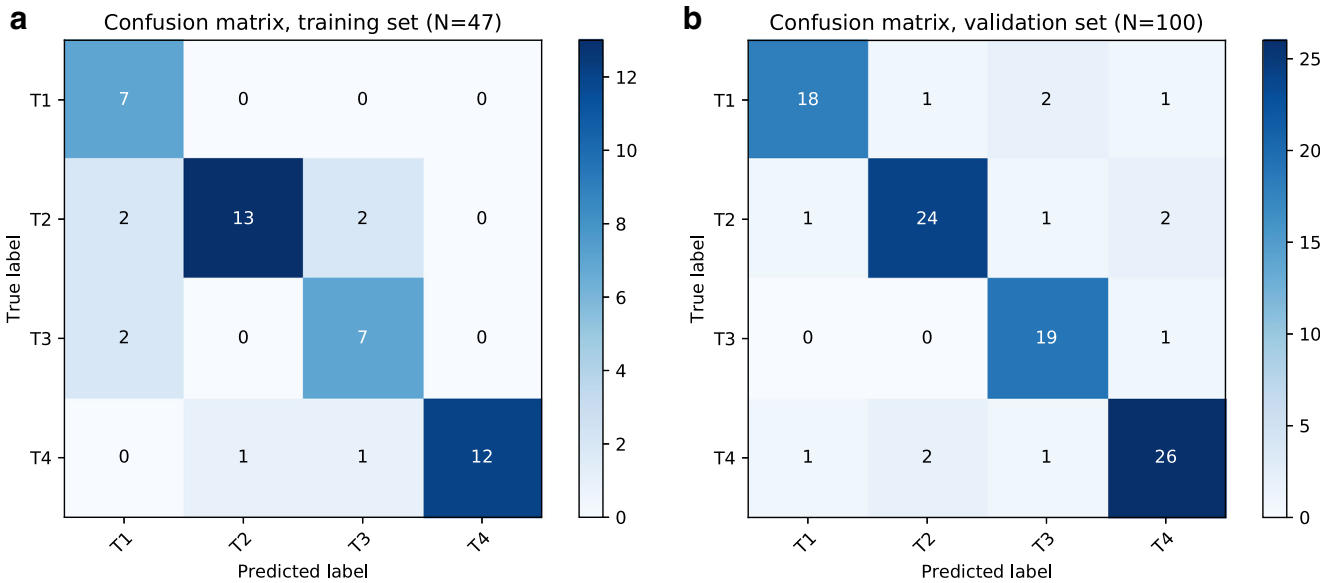

Fig. 2 Confusion matrices of T-stage classification Confusion matrices of the T-stage classification on the training and validation sets

Table 4 T-stage errors by category

Error group	Error type	Description	Training ($n = 47$)	Validation ($n = 100$)
Context	Context missing	Context not matched because of missing modifier	0	1
	Context mismatch	Context mismatch, wrong modifier detected	2	3
	Context disagreement	Disagreement about certain/prob. certain	0	1
Concepts	Missing synonym	Concept not matched because of a missing synonym or expression.	2	0
	Algorithm logic	Presence or involvement not correctly classified	0	2
Standardization	Measurement extractor	e.g., using expressions (more than 5 cm) or 4–51 op 11 cm, blacklist for size	2	2
	Dictation artifact	Errors related to dictation (e.g., whitespaces within numbers)	0	1
	Standardization	Wrong heading above section	0	1
Complexity	T4 multiple lobes	Error related by detecting tumor present in multiple lobes of the same long	1	1
spaCy	Sentence Boundary Detection	Error in detecting the boundary of a sentence, therefore involvement logic does not hold	0	1
	Total errors		7	13

T-stage errors by category for the training and validation sets

Lexicon (RadLex) was not available in Dutch and could therefore not be tested. Ideally, a standardized vocabulary should be used to standardize data and try to make data more uniform. Data should then be labeled with SNOMED-CT or RadLex codes in order to increase findability, according to the Findable, Accessible, Interoperable, and Reusable (FAIR) principles [18].

Another important finding is that radiological free text reports consist of many contextual expressions, phrases, and words (see Appendices 2 and 3), which are indispensable for accurate description of a specific disease. For instance, concepts should be properly correlated to the right context like negations or sizes, but the same holds for probabilities and the extent of involvement. This is a difficult and important process and should be done with care, because context allows radiologists to nuance and specify their findings. This lack of nuancing possibilities is probably one of the caveats of structured reporting and its broad implementation.

When analyzing the errors in detail, one can see that the errors made are diverse, although most wrongly staged tumors were related to context extraction (35%). Several times there is a mismatch between concept and context caused by the shallow approach of pyContextNLP. For example, when two concepts are present in the same sentence, context (e.g., a negation) can be matched with the wrong concept. This might be overcome by dependency parsing which can improve contextual matching.

This paper tried to divide pre-processing and processing steps in order to differentiate errors found, but the errors are often hard to separate, as both steps are highly intertwined. For instance, errors made by the sentence splitter can be related to the fact that the model is not trained on medical reports. However, errors can also be introduced by radiological reporters using a different (staccato) way of reporting. The use of speech recognition in radiological reporting introduces several imperfections, mainly resulting in incorrect punctuation and white space errors within numbers. This can only be partly improved by pre-processing steps.

Task complexity is a different hurdle to overcome. Problems might, for instance, arise when concepts of different items should be combined in a single statement (e.g., T4-stage, different lobes, same lung) or should be ignored (e.g., gravity depending atelectasis vs. tumor related atelectasis). This is especially the case when these concepts are stated in different sentences. Specific annotation guidelines or agreements can partly improve this difficulty. However, algorithms should not be unnecessarily more complicated when steps like standardization of the report content or reporting manner can increase report homogeneity. This is highlighted by the errors made in the standardization category (30%) which is related to the input of the reporter and dictation technology used. Standardizing report content by using a certain standardized language, for instance, the vocabulary used in the TNM classification, will result in less synonyms in the report. In addition, when sentences stated are less ambiguous, by for instance, stating only information about the described item in the same sentence, outcomes will further be improved. As such, standardization of reporting content and manner will improve outcomes without expanding existing algorithms. Hence, NLP and standardization are counterparts in which high-end NLP tooling makes standardization redundant, but proper standardization can improve the structured data and the accuracy of the NLP tool.

Several limitations of this study should be mentioned of which the small sample size is the most important one. Furthermore, this algorithm is only trained at one specific

dataset of one radiological department. Therefore, overfitting is a concern. Although this has not been the main goal of this pilot, future work should focus on external validation.

In addition, future work should be done to explore how NLP algorithms can increase the value of the radiological report when, for instance, they are incorporated in the reporting process. Live classifications can be displayed when an algorithm is processing the free text during reporting. An algorithm can also notify the reporter when information about a specific item is missing. In addition, this tumor staging algorithm can also be used for restaging earlier staged tumors according to the current TNM edition. As such, NLP algorithms can be used in various ways to enhance reporting content and support the FAIR principles.

Conclusion

NLP is a promising technology for mining free text radiological reports and can be introduced in English and in a small, non-English language such as Dutch. However, the proper implementation of a free text algorithm depends largely on the context of concepts mentioned in the report, more than on specific words. Implementing NLP and standardization should be balanced, and ratios adjusted depending on the available data. Future work should mainly focus on how to (gradually) use more machine or deep learning approaches.

References

1. McGinty GB, Allen B, Geis JR, Wald C: IT infrastructure in the era of imaging 3.0. J Am Coll Radiol 11:1197–1204, 2014
2. Brierley J, Gospodarowicz MK, Wittekind C Eds: TNM classification of malignant tumours, 8th edition. Chichester: John Wiley & Sons Inc., 2017
3. Puts S, Nobel JM: Medical narrative to structure: maastroclinic/medstruct. maastroclinic, 2019
4. Krupinski EA, Hall ET, Jaw S, Reiner B, Siegel E: Influence of radiology report format on reading time and comprehension. J Digit Imaging 25:63–69, 2012
5. Pons E, Braun LMM, Hunink MGM, Kors JA: Natural language processing in radiology: A systematic review. Radiology 279:329–343, 2016
6. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG: Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. J Am Med Inform Assoc 17:507–513, 2010
7. Cornet R, van Eldik A, de Keizer N: Inventory of tools for Dutch clinical language processing. Stud Health Technol Inform 180:245–249, 2012
8. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, Colquist S: Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc 17:440–445, 2010
9. Castro SM, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S, Bekhuis T, Jacobson RT: Automated annotation and classification of BI-RADS assessment from radiology reports. J Biomed Inform 69:177–187, 2017
10. Pathak S, van Rossen J, Vijlbrief O, Geerdink J, Seifert C, van Keulen M: Automatic Structuring of Breast Cancer Radiology Reports for Quality Assurance. IEEE international conference on data mining workshops (ICDMW), Singapore, IEEE 2018(732–739):2018, 2018
11. Honnibal M, Montani I: Spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear: 7, 2017
12. Soldaini L, Goharian N: QuickUMLS: a fast, unsupervised approach for medical concept extraction. MedIR workshop, sigir, 2016. **Available at** http://ir.cs.georgetown.edu/downloads/quickumls.pdf. Accessed 6 May 2019.
13. Côté RA, Robboy S: Progress in medical information management. Systematized nomenclature of medicine (SNOMED). JAMA 243:756–762, 1980
14. Chapman BE, Lee S, Kang HP, Chapman WW: Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform 44:728–737, 2011
15. Chapman WW, Hillert D, Velupillai S, Kvist M, Skeppstedt M, Chapman BE, Conway M, Tharp M, Mowery DL, Deleger L: Extending the NegEx lexicon for multiple languages. Stud Health Technol Inform 192:677–681, 2013
16. Afzal Z, Pons E, Kang N, Sturkenboom MC, Schuemie MJ, Kors JA: ContextD: An algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. BMC Bioinformatics 15:373, 2014
17. Chapman WW: Extract context modifiers targeting clinical terms: Maastroclinic/pyConTextNLP 2019. **Available at** https://github.com/maastroclinic/pyConTextNLP. Accessed 19 June 2019.
18. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018, 2016