Machine learning to parse breast pathology reports in Chinese

As Published: 10.1007/S10549-018-4668-3

Publisher: Springer Science and Business Media LLC

Persistent URL: https://hdl.handle.net/1721.1/134872

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Breast Cancer Research and Treatment
https://doi.org/10.1007/s10549-018-4668-3

PRECLINICAL STUDY

Machine learning to parse breast pathology reports in Chinese

Rong Tang[1] · Lizhi Ouyang[2] · Clara Li[3] · Yue He[2] · Molly Griffin[1] · Alphonse Taghian[4] · Barbara Smith[1] · Adam Yala[3] · Regina Barzilay[3] · Kevin Hughes[1]

Received: 8 January 2018 / Accepted: 11 January 2018

Abstract

Introduction Large structured databases of pathology findings are valuable in deriving new clinical insights. However, they are labor intensive to create and generally require manual annotation. There has been some work in the bioinformatics community to support automating this work via machine learning in English. Our contribution is to provide an automated approach to construct such structured databases in Chinese, and to set the stage for extraction from other languages.

Methods We collected 2104 de-identified Chinese benign and malignant breast pathology reports from Hunan Cancer Hospital. Physicians with native Chinese proficiency reviewed the reports and annotated a variety of binary and numerical pathologic entities. After excluding 78 cases with a bilateral lesion in the same report, 1216 cases were used as a training set for the algorithm, which was then refined by 405 development cases. The Natural language processing algorithm was tested by using the remaining 405 cases to evaluate the machine learning outcome. The model was used to extract 13 binary entities and 8 numerical entities.

Results When compared to physicians with native Chinese proficiency, the model showed a per-entity accuracy from 91 to 100% for all common diagnoses on the test set. The overall accuracy of binary entities was 98% and of numerical entities was 95%. In a per-report evaluation for binary entities with more than 100 training cases, 85% of all the testing reports were completely correct and 11% had an error in 1 out of 22 entities.

Conclusion We have demonstrated that Chinese breast pathology reports can be automatically parsed into structured data using standard machine learning approaches. The results of our study demonstrate that techniques effective in parsing English reports can be scaled to other languages.

Keywords Machine learning · Natural language processing (NLP) · Pathology reports · Electronic health record (EHR) · Chinese

Introduction

Natural language processing (NLP) has been used extensively in industry and government to turn free text into structured, machine readable data; however, its use in medicine has been limited.

While hospitals contain vast amounts of clinical data, a small fraction of it is structured and used to derive insights, which is in stark contrast to other industries. Almost all critical medical information is recorded in the Electronic Health Record (EHR) as free text. The use of this data has, to date, required manual curation to turn free text into structured data. Manual extraction of information from the EHR is time-consuming and specialized; it requires efforts by well-educated individuals, often at the MD level. The extreme difficulty and high cost of manual extraction limits the size of data sets.

Increasing numbers of researchers are attempting to use NLP to decipher the information locked in the EHR. There were a few studies applied NLP in other fields [1–3]; however, we were only able to find one published brief report of Chinese medical NLP [4]. The authors used an Excel VBA rules-based NLP system that extracted 15 entities

✉ Kevin Hughes

Rong Tang
rtang1@partners.org

1 Division of Surgical Oncology, MGH, Boston, USA

2 Department of Breast Surgery, Hunan Cancer Hospital, Changsha, Hunan, China

3 Department of Electrical Engineering and Computer Science, CSAIL, MIT, Cambridge, USA

4 Department of Radiation Oncology, MGH, Boston, USA

Published online: 29 January 2018

from breast cancer pathology reports and demonstrated higher speed and comparable accuracy when compared to a Chinese Tumor Registrar. Although this small study demonstrated the feasibility of NLP of Chinese medical records in 2006, there has been no follow-up publication. The literature does not mention any other systems.

The main approaches to NLP are rules based and machine learning. In the rules-based approach, all words and phrases that denote a given entity are cataloged and their presence, absence, or negation in a given document is identified. In the machine learning approach, the computer is given examples of the presence, absence, or negation of a given entity and the machine learns to map patterns in the free text to the entity labels.

The inherent linguistic and structural variability of free text poses challenges to using the rules-based approach for efficient data retrieval. Additionally, any use of the rules-based approach would require large amounts of expert engineering to achieve sufficient performance; for example, knowing the 124 ways of saying "invasive ductal carcinoma" in English [5]. With this study, we were about to turn the power of machine learning to parsing Chinese pathology reports into machine readable data.

Methods

Pathology reports retrieval and entities selection

With the approval of the Partners Institutional Review Board (IRB) and Hunan Cancer Hospital IRB, all electronically available pathology reports from the Breast Surgery Department of the Hunan Cancer Hospital from January 1, 2016 to August 31, 2016 were retrieved. A total of 2104 breast pathology reports in Chinese were collected. The pathology reports differed widely in structure and verbalization (word use/terminology/syntax) between individual pathologists. We isolated the diagnostic part of each pathology report for use in this study and removed all protected health information, medical histories, and gross descriptions from the reports. Patient identification numbers were recorded as a code with keys only available in the Hunan Cancer Hospital.

Using the College of American Pathologists' (CAP) synoptic reporting system and expert opinion, we identified 29 types of information (Entities) that could be found in pathology reports and were known to be useful in categorizing breast disease and breast cancer (Table 1). Many of the same entities had been selected and analyzed in our previous study of English pathology reports [5, 6]. Entities fell into two categories: binary (yes/no, left/right, and positive/negative) and numeric. 29 entities were initially collected in the study; however, we eliminated 7 entities due to a lack of annotated training examples.

Table 1 Entities collected and annotated in this study

Entities from CAP synoptic reporting system	Type of data
Laterality (left/right)	Binary
Invasive ductal carcinoma (P/A)	Binary
Invasive lobular carcinoma (P/A)[a]	Binary
Ductal carcinoma in situ (P/A)	Binary
Any cancer (P/A)	Binary
Lobular carcinoma in situ (P/A)[a]	Binary
Atypia (P/A)	Binary
Positive lymph nodes (#)	Numerical
Estrogen receptor status (Pos/Neg)	Binary
Progesterone receptor status (Pos/Neg)	Binary
HER 2 status (Pos/Neg)	Binary
Other common pathologic diagnoses in Hunan Cancer Hospital	
Specimen from biopsy (Y/N)	Binary
Tumor size (#)	Numerical
Tumor grade (#)	Numerical
Cancer suspicious (P/A)[a]	Binary
Invasive cancer NOS (P/A)	Binary
Any Invasive cancer (P/A)	Binary
Invasive suspicious (P/A)[a]	Binary
Cancer NOS (P/A)[a]	Binary
% Estrogen receptor (#)	Numerical
% Progesterone receptor (#)	Numerical
% KI-67 (#)	Numerical
Nipple/skin involvement (P/A)[a]	Binary
Margin status (Pos/Neg)[a]	Binary
Lymph nodes removed (#)	Numerical
Hyperplasia without atypia (P/A)	Binary
Fibroadenoma (P/A)	Binary
Papillary lesion (P/A)	Binary
Phyllodes (P/A)[a]	Binary

Binary entities: P/A: present/absent, Pos/Neg: positive/negative Y/N: yes/no

Numerical entities: #: number

NOS not otherwise specified

[a]These eight entries were collected but were found in fewer than 75 binary cases after annotation. Accordingly, they were excluded from further study

Database creation and pathology reposts annotation

The work-flow for processing these reports was as follows:

First, the Chinese text of the diagnostic section of each pathology report was copied into a cell in an Excel spreadsheet, with each case/side occupying a single line. We included bilateral cases (with the same MRN code) when

there were separated reports for each sides and excluded the 78 with bilateral breast information in the same report. That left us 2026 pathology reports for further study.

Second, we created a Database table with a field for each entity is listed in Table 1. "Atypia" in this study was a generic group of any ADH, ALH, and atypical hyperplasia.

"Specimen from biopsy" entities were marked "yes" if a specimen was from a core or excisional biopsy. "Invasive cancer NOS" included a diagnosis of invasive cancer which did not specify invasive ductal, invasive lobular, or other specific type. Similarly, "Cancer NOS" included those with cancer that did not specify invasive versus non-invasive. "Invasive suspicious" was marked present when DCIS was present but unclear if invasion was present.

Third, physicians who were fluent in Chinese and English and familiar with breast pathology, extracted the entities from each report. They recorded the values of each entity. Not all entities were available for each case. For example, ER status and PR status were only available for cancer cases, and for these entities, we only evaluate our algorithm on cancer cases.

Finally, the de-identified dataset was sent to the US through secure email.

Machine learning method

To group the Chinese characters in the text into words, we used Jieba [7], a commonly used Chinese word segmentation tool, on all reports. The 2026 segmented report texts and their corresponding MD annotations were divided into three sets: 1216 (60%) for training, 405 (20%) for development, and 405 (20%) for testing.

We used the training set to train an independent model for each entity. For binary entities like DCIS, with possible labels of "Present" and "Absent", we developed a boosted classifier. For numerical entities for which the correct label was an exact string in the report text, such as tumor size or ER stain percentage where we needed to find a specific number, we developed a conditional random field tagger. Both methods are discussed below.

Boosted classifier for binary entities

We trained a model to parse binary entities using boosting classification. In boosting classifiers, strong non-linear classification is achieved by combining weak learners, such as decision stumps. We represented each pathology report using a standard n-gram representation, where the text was mapped to a vector capturing words and phrases that appeared within it. During training, the model learned the weights of each phrase. We trained the classifier separately for each one of the binary entities.

Conditional random field tagger for numerical entities

For entities where the annotation takes the form of a string inside the text, we developed a conditional random field [8] tagger. For instance, given the report text "The ER percentage value was 80%", the classifier has to identify that "80%" is the ER stain percentage. In this setting, we trained the model to predict whether a given word in its context, was the correct value for that entity. During training, the model learned how to weigh each word, and how to weigh the possible contexts in which it appeared to predict the labels. A diagram of the algorithm is as depicted in Fig. 1).

Performance evaluation

Each entity was studied independently.

For each entity, the following measures were calculated: Precision, Recall, F1 scores, and Accuracy scores.

Accuracy is the overall accuracy rate, what percentage of predicted labels matched the annotated labels. Precision is the positive predictive value, out of all those cases that the machine labeled as positive, how many were actually positive. Recall is the true positive rate, also called sensitivity in binary classification, where out of all the cases that were annotated as positive, how many did the machine correctly predict as positive. F1 score is the harmonic mean of precision and sensitivity. We evaluated our performance by comparing our model's predictions on a held-out set of 405 pathology reports for each entity against their corresponding MD annotations.

To evaluate per-report performance, we computed the ''all-or-nothing'' report-level accuracy where the system must correctly extract all the information about the 21 entities from a given report to be correct.

We studied the overall performance, which is an average of individual entity scores and performed a Learning Curve Analysis where we plotted the performance of the system, measured in aggregate F score over entities, as

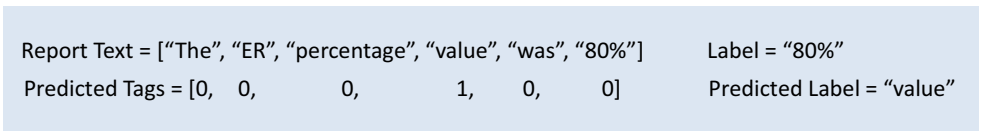

Fig. 1 A diagram of the tagging task algorithm

we varied the amount of annotation from 10 examples to 1000 examples. This type of analysis let us analyze how much annotation of our model needed to start achieving reasonable performance.

Results

Overall, 2104 diagnostic part of de-identified pathology reports were collected from Hunan Cancer Hospital, 78 reports involving bilateral results that could not be separated were excluded. The remaining 2026 reports included 1198 reports with cancer, 29 reports suspicious for cancer, and 799 reports with only benign disease.

Extraction Results

An example of system input and output is shown in Table 2. The average length of the diagnostic part of a pathology report was 88 words.

Per-entity performance

Table 3 demonstrates the performance of 13 binary entities by using a bag-of-words model, and Table 4 shows the performance of the eight numerical entities by applying a tagging task model.

In Table 3, most of the entities had 1216 training samples. However, ER and PR status had fewer cases as only non-cancer cases did not have ER or PR reposted, and majority but not all cancer cases had these entities available (86 of the 719 cancer cases in the training set did not have ER and PR reported, leaving 633 reports with results). This was also true for the test set and the development set.

Table 3 also shows the performance evaluation of the binary entities for the test set. 12 of the 13 entities had an entity-average F score over 0.95, and overall accuracy greater than 95%. The PR status is the only entity with an F score 0.91 and an overall accuracy of 91%. However, in the present category of "invasive cancer NOS" and "atypia" (marked in bold), the variate-specific F score, precision, and recall were lower than 0.9.

Table 2 A sample of pathology report in Chinese showing the fields extracted (highlighted in bold type) and extraction results

A de-identified pathology report in Chinese	Extracted diagnostic information from the sample pathology report	
姓名：*	Laterality	Left
年龄：*	Specimen from biopsy	No
住院号：NHUOJE **	IDC	Present
手术日期：*	Invasive cancer NOS	Absent
（镜下表现截图）*	Any Invasive cancer	Present
	DCIS	Present
1.（左乳改良根治术标本）浸润性导管癌Ⅱ-Ⅲ级，部分为导管内癌，肿块大小约 **2.5*2*2cm**，脉管内未见明确癌栓；	Any cancer	Present
	Tumor Size (cm)	2.5
2.（左腋下）**淋巴结：11/15** 见癌转移；	Tumor Grade	2-3
3. **皮肤、基底切缘、乳头均未见癌累及**。	Positive lymph nodes	11
	Lymph nodes removed	15
ER: + (约 70%)	ER status	Positive
PR: 2+ (约 80%)	ER %	70%
CerbB-2: - (0)	PR status	Positive
Ki-67: 约 70%	PR %	80%
CK5/6: 小灶+	HER-2 status	negative
P63: 小灶+	Atypia	Absent
EGFR: +	Hyperplasia without atypia	Absent
P53: -	Fibroadenoma	Absent
	Papillary lesion	Absent
病理医生签名：**	% KI-67	70%

Note English notations for ER, PR and some other entities are used in Chinese reports

* This information, including name and age of the patient, date of surgery, image of typical lesion, and signature of the pathologist was in the original report but not collected in this study

** MRN is code protected by Chinese hospitals; code is unique for each patient

Table 3 Extraction accuracy, precision, recall, and *F* scores for the 13 binary entities

Entities	Value	Training	Testing results (sample size = 405[a])			
		Sample size	Precision	Recall	F score	Accuracy
Breast side	Right	583	0.98	0.98	0.98	0.98
	Left	633	0.98	0.98	0.98	
	Total/avg	1216	0.98	0.98	0.98	
Specimen from biopsy	Yes	303	0.92	1.00	0.96	0.98
	No	913	1.00	0.98	0.99	
	Total/avg	1216	0.98	0.98	0.98	
IDC	Present	475	0.96	0.98	0.95	0.96
	Absent	741	0.96	0.93	0.97	
	Total/avg	1216	0.96	0.96	0.96	
Invasive cancer NOS	Present	**167**	**0.85**	**0.85**	**0.85**	0.97
	Absent	1049	0.98	0.98	0.98	
	Total/Avg	1216	0.97	0.97	0.97	
Any invasive cancer	Present	653	0.95	0.96	0.95	0.96
	Absent	563	0.97	0.95	0.96	
	Total/Avg	1216	0.96	0.96	0.96	
DCIS	Present	145	0.96	0.95	0.96	0.99
	Absent	1071	0.99	0.99	0.99	
	Total/Avg	1216	0.99	0.99	0.99	
Any cancer	Present	719	0.98	0.96	0.97	0.98
	Absent	497	0.98	0.99	0.98	
	Total/Avg	1216	0.98	0.98	0.98	
Atypia	Present	**80**	**1.00**	**0.68**	**0.81**	0.98
	Absent	1136	0.98	1.00	0.99	
	Total/Avg	1216	0.98	0.98	0.98	
Fibroadenoma	Present	228	1.00	0.97	0.99	1.00
	Absent	988	0.99	1.00	1.00	
	Total/Avg	1216	1.00	1.00	1.00	
Hyperplasia without atypia	Present	352	0.93	0.95	0.94	0.97
	Absent	864	0.98	0.97	0.98	
	Total/Avg	1216	0.97	0.97	0.97	
Papillary lesion	Present	99	0.96	1.00	0.98	1.00
	Absent	1117	1.00	1.00	1.00	
	Total/Avg	1216	1.00	1.00	1.00	
ER status	Positive	423	1.00	0.95	0.97	0.97
	Negative	210	0.92	1.00	0.96	
	Total/Avg	633	0.97	0.97	0.97	
PR status	Positive	423	0.97	0.89	0.93	0.91
	Negative	210	0.82	0.95	0.88	
	Total/Avg	633	0.92	0.91	0.91	

The average precision, recall, and *F* scores displayed in the "total" row are weighted by the number of cases for each value

Total/avg Total case number or average entity score

[a]The sample size for the testing set was 405 for all the entities, except ER status and PR status

In Table 4, all the average F scores were above 0.97, and all the average accuracies were greater than 0.93, except "# removed nodes", which had an accuracy of 0.91. Since some reports contained both the number of sentinel lymph nodes and the number of axillary dissection harvested lymph nodes, the machine would have needed to add these two numbers to arrive at the correct answer, contributing to the lower accuracy score. This step requires more training or processing than just identifying one number in the reports.

Table 4 Extraction accuracy, precision, recall, and F scores for the 8 numerical entities

Entities	Training sample size[a]	Testing results (testing sample size varies)[a]			
		Avg. precision	Avg. recall	Avg. *F* score	Avg. accuracy
Size of main lesion	459	1.00	0.98	0.99	0.93
Grade	415	1.00	1.00	1.00	1.00
ER stained	437	1.00	0.98	0.99	0.95
PR stained	433	1.00	0.99	0.99	0.95
Her2 receptor	624	1.00	0.98	0.99	0.96
Ki67	634	1.00	0.99	1.00	0.96
# Pos nodes	1216	0.99	0.99	0.99	0.97
# Removed nodes	1216	0.99	0.95	0.97	0.91

[a]The training and testing sample size varies, because we only included cases with the information for these entities available

Table 5 Overall performance as an average of individual entity scores

Entities	Precision	Recall	F score	Accuracy
Binary entities	0.98	0.98	0.98	0.98
Numerical entities (all)	1.00	0.98	0.99	0.95
Total (all)	0.98	0.98	0.98	0.97

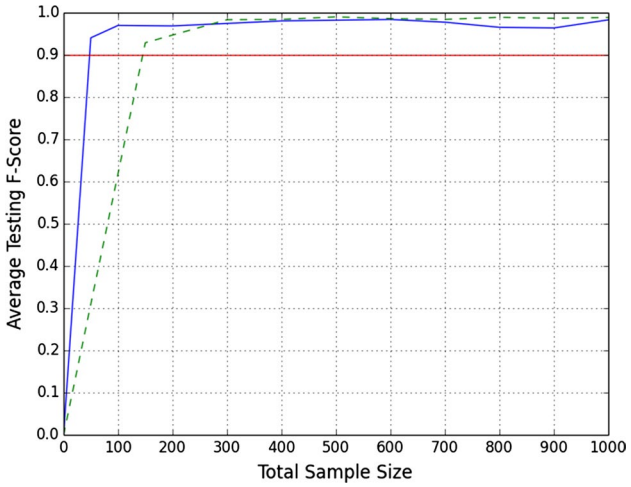

Fig. 2 Learning curve. The total sample size (including training, development, and testing) plotted against the average testing *F* score (blue, solid: average *F* score for all binary entities; green, dash: average *F* score for all numerical entities)

Per-report performance

For binary entities (excluding the entities with NA as a possible value), the system achieved completely correct per-report extraction for 346 out of 405 reports (85%), 42 (11%) had one error and 17 (4%) had more than one error.

Overall performance

Table 5 demonstrates the overall performance as an average of individual entity scores.

Learning curve analysis

Figure 2 shows the learning curve analysis. We found that for binary entities, F scores reached above 90% with 50 total (training, development, and testing) cases, and about 98% with 500 cases. For numerical entities, F score reached above 90% with about 150 total cases, and about 98% with 500 cases.

Discussion

A recently published review by Burger et al. [9] examined 38 papers that used natural language processing (NLP) to extract and encode clinically relevant information from pathology reports. Rule-based methods were used with varying degrees of efficacy [5, 10–14]. We [5] understand the frustrations of using this approach, having found that the diagnosis of invasive ductal carcinoma (IDC) had been recorded in 124 different ways in our health care system with 13 possible ways to negate the finding. Similar findings have been reported by others who employed rules-based approaches for NLP [5, 10–14].

Alternatively, machine learning methods relied on an automatic capacity for learning extraction patterns from a set of manually annotated pathology reports (the training set) and for then extrapolating those patterns to new reports. The accuracy of the algorithm depends on the size of the training set, the complexity of the task, and the strength of the learning algorithm.

We chose to study pathology reports for this study as they are essential to cancer registry systems, identification of patients for clinical trials and other research and quality

improvement efforts, but almost always exist only as free text. NLP has the potential to reduce the effort required to analyze large amounts of data from medical records, which would reduce the cost and time required to glean scientific insight from this data. A number of researchers have employed machine learning techniques for parsing breast pathology reports [15–18] with varying success.

We recently described [6], a machine learning approach to NLP for automatically extracting pertinent entities from English pathology reports. We trained our system with annotations from two datasets, consisting of 6295 and 10,841 manually annotated reports. The model achieved a by-report accuracy of 90% for correctly parsing all carcinoma and atypia entities for a given patient. The average accuracy for individual entities was as high as 97%. In that study, a database of 91,505 parsed English pathology reports was created, which has since been enlarged and enhanced.

With this report, we document the first attempt to apply a machine learning NLP approach to Chinese medical records. Our results demonstrate that this approach can be used to extract information from Chinese breast pathology reports. Compared to our previous study using English pathology reports [6], the training set volume is less than 10% of that used in our previous study, with only 1216 cases. In addition, we expanded the study to both binary and numerical entities. Even with less training and more types of entities, the average per-report accuracy for correctly parsing all binary entities for this study is 85% and overall per-entity accuracy of binary and numerical entities is as high as 97%.

There are several possible reasons for this finding. In our prior study [6], the average length of each pathology report after preprocessing was 354 English words, including the identifying data, the history of the disease, the diagnostic description, and the gross description. In this study, the average length of each Chinese pathology report was 88 words, including only the diagnostic description. Second, there are more ways to express the same entity in English than in Chinese. For example, our group [5] showed that "invasive ductal carcinoma" (23 letters) had been stated in 124 different ways in their English pathology reports, including abbreviations and typos. Alternatively, in Chinese, there are very few ways to say invasive ductal carcinoma, with the concept usually expressed as 5 characters, "浸润性导管癌", with little likelihood of abbreviation or misspelling.

In Table 3, three entities—the precision, recall, and *F* score of present/positive—were lower than 0.9 (marked in bold) because the number of reports with present/positive in these entities was low in both the training set and the testing set. This lower performance is to be expected with insufficient training examples and may be resolved with a bigger dataset.

We believe this approach will widely empower clinicians to utilize information locked in EHR in the future. Our next step would be collecting a set of pathology reports from another Chinese institution to test if our model, trained on one institution, is able to generalize across institutions. In the meantime, we plan to apply this machine learning approach to natural language processing to extract clinical information from pathology reports in other languages to see if this approach can be generalized.

References

1. Huang CR, Chen KJ, Chang LL (1996) Segmentation standard for Chinese natural language processing. In: Proceedings of the 16th conference on Computational linguistics, vol. 2 (pp. 1045–1048). Association for Computational Linguistics
2. Wong KF, Li W, Xu R, Zhang ZS (2009) Introduction to Chinese natural language processing. Synth Lect Hum Lang Technol 2(1):1–148
3. Qiu X, Qi Z, Huang X (2013) Fudan NLP: a toolkit for Chinese natural language processing. In: ACL (conference system demonstrations), pp. 49–54
4. Liang YF, Chu PY, Chang CS, Wang CH, Chang P (2006) Developing and evaluating a simple, spreadsheet-based pathology report extraction system for cancer registrars. AMIA Ann Sym Proc 2006:1008
5. Buckley JM, Coopey SB, Sharko J, Polubriaginof F, Drohan B, Belli AK, Kim EM, Garber JE, Smith BL, Gadd MA et al (2012) The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 3:23
6. Yala Adam, Barzilay Regina, Salama Laura, Griffin Molly, Sollender Grace, Bardia Aditya, Lehman Constance et al (2017) Using machine learning to parse breast pathology reports. Breast Cancer Res Treat 161(2):203–211
7. Sun J (2013) Jieba (version 0.39) [source code]. https://github.com/fxsjy/jieba
8. Korobov M (2015) Sklearn-crfsuite (Version 0.3.6) [source code] https://github.com/TeamHG-Memex/sklearn-crfsuite
9. Burger G, Abu-Hanna A, de Keizer N, Cornet R (2016) Natural language processing in pathology: a scoping review. J Clin Pathol 69(11):949–955
10. Edwards GA (2008) Expert systems for clinical pathology reporting. Clin Biochem Rev 29:S105–S109
11. Napolitano G, Fox C, Middleton R, Connolly D (2010) Pattern based information extraction from pathology reports for cancer registration. Cancer Causes Control 21:1887–1894
12. Nguyen A, Lawley M, Hansen D, Colquist S (2011) Structured pathology reporting for cancer from free text: lung cancer case study. Electron J Health Inform 7:8
13. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, Colquist S (2010) Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc 17:440–445
14. Weegar R, Dalianis H (2015) Creating a rule based system for text mining of Norwegian breast cancer pathology reports. In: Sixth international workshop on health text mining and information analysis (Louhi), p 73
15. Li Y, Martinez D (2010) Information extraction of multiple entities from pathology reports. In: Australasian Language Technology Association Workshop, p 41
16. Martinez D, Li Y (2011) Information extraction from pathology reports in a hospital setting. In: Proceedings of the 20th ACM

international conference on information and knowledge management, ACM, pp 1877–1882
17. Nguyen A, Moore D, McCowan I, Courage M-J (2007) Multiclass classification of cancer stages from free-text histology reports using support vector machines. In: 29th annual international conference of the IEEE engineering in medicine and biology society, IEEE, pp 5140–5143
18. Wieneke AE, Bowles EJ, Cronkite D, Wernli KJ, Gao H, Carrell D, Buist DS (2015) Validation of natural language processing to extract breast cancer pathology procedures and results. J Pathol Inform 6:38