Natural Language Processing for Automated Extraction of Breast Cancer Information for the Cancer Registry

Adhari Abdullah Alzaabi, MD, PhD[1] and Abdulrahman AAlAbdulsalam, PhD[2]
[1] College of Health Science and Medicine, Sultan Qaboos University [2] Department of Computer Science, Sultan Qaboos University

Background. National cancer registries rely on manual abstraction of free-text clinical records to collect vital information about cancer diagnosis, stage, progression and treatment [1]. Many prior studies have demonstrated the ability of natural language processing (NLP) based on machine learning to extract information from free-text clinical records for a variety of purposes (diagnosis, adverse events discovery, clinical trial matching, …, etc.) [2]. We present in this study experimental results of applying NLP to extract information from the records of breast cancer patients for the cancer registry in Oman.

Methods. After obtaining ethical approval from two local institutions (Sultan Qaboos University Hospital and Royal Hospital), the clinical records (pathology, oncology and surgical notes) were collected for 1152 patients (462 from SQUH and 690 from Royal) who have been diagnosed with breast cancer in the years 2013 to 2018. Manually abstracted data within the cancer registry databases for the same patients were extracted to serve as the gold standard to evaluate the NLP approaches. We experimented with two approaches for information extraction from free-text clinical records: 1) using the readily available **DeepPhe** system [3], and 2) rule-based regular expression matching approach (**REGEX**). The precision (positive predictive value), recall (sensitivity) and F1 metrics were used to report the performance of each approach.

Results. The table below shows the performance of each approach on data from each institution (SQUH and Royal) on six categories of information: Cancer primary site, histology, grade, pathological TNM stage, and summary stage. The DeepPhe system was able to identify the primary site of cancer for the majority of cases from Royal (F1=0.91), however, it scored quite low (F1=0.47) for cases from SQUH. On the other hand, the REGEX system was almost perfect in identifying the cancer site for all cases (F1=0.99/1.0). For the breast cancer histological type, laterality and grade, the REGEX system performed better than DeepPhe especially for the cases from SQUH with overall difference in F1 ranging from +0.04 (for grade in Royal data) to +0.63 (for histology in SQUH data). The low performance of DeepPhe especially for histological types of cancer in SQUH data was due the fact that the system did not generate any output for many patient cases (recall=0.12 vs. prec.=0.85). Extraction of TNM and summary staging was most challenging for both systems (F1 ranges of 0.18 to 0.53). We suspect the reason is due to the lack of proper documentation of the TNM stage in free-text clinical records for many cases. The REGEX system had the advantage over DeepPhe with the addition of the rule to predict TxNxMx by default when no matching TNM phrase is found in the patient records (recall for REGEX is much higher than DeepPhe for staging information).

		Primary Site		**Histological type of Cancer**		**Laterality**		**Grade**		**T**		**N**		**M**		**Summary (Stage)**	
		Royal	**SQUH**	**Royal**	**SQUH**	**Royal**	**SQUH**	**Royal**	**SQUH**	**Royal**	**SQUH**	**Royal**	**SQUH**	**Royal**	**SQUH**	**Royal**	**SQUH**
REGEX	**Prec.**	1	1	0.87	0.86	0.83	0.76	0.52	0.65	0.54	0.59	0.59	0.61	0.57	0.6	0.46	0.32
	Recall	0.97	0.99	0.69	0.82	0.69	0.83	0.52	0.57	0.25	0.51	0.26	0.53	0.17	0.51	0.63	0.41
	F1	0.99	1	0.76	0.84	0.75	0.79	0.51	0.53	0.18	0.45	0.2	0.47	0.22	0.53	0.53	0.36
DeepPhe	**Prec.**	1	1	0.89	0.85	0.82	0.74	0.54	0.34	0.33	0.38	0.69	0.56	0.61	0.54	0.5	0.31
	Recall	0.83	0.3	0.47	0.12	0.62	0.25	0.47	0.14	0.13	0.09	0.14	0.08	0.1	0.07	0.67	0.55
	F1	0.91	0.47	0.61	0.21	0.7	0.37	0.47	0.16	0.13	0.14	0.14	0.13	0.13	0.12	0.55	0.4

Conclusion. The study presented our attempt to use an existing NLP system (DeepPhe) and rule-based NLP system (REGEX) developed inhouse for automatic extraction of information from breast cancer patient records for the cancer registry. The DeepPhe system demonstrated the feasibility of detecting primary site information and moderate performance for histology and laterality information for data from one institution (Royal). However, DeepPhe performance was very low on data from another institution (SQUH). One possible explanation is that DeepPhe was trained on radiology and surgical pathology reports with sections structured differently from pathology reports of SQUH. Machine learning models are well known to be very sensitive to changes in datasets that were used during training. The REGEX system demonstrated feasibility for extraction of primary site, histology and laterality of breast cancer (F1 >= 0.75), however, the performance on the rest of information categories (grade, and stage) was quite low (F1 <= 0.53). One limitation is that the current evaluation was conducted at the patient-level. In future work, we plan to create manually labeled reference dataset to be used for training and evaluation of NLP models at the document level.

References

1. AAlAbdulsalam, Abdulrahman K., et al. "Automated extraction and classification of cancer stage mentions fromunstructured text fields in a central cancer registry." *AMIA Summits on Translational Science Proceedings* 2018 (2018): 16.
2. Kreimeyer, Kory, et al. "Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review." *Journal of biomedical informatics* 73 (2017): 14-29.
3. Savova, Guergana K., et al. "DeepPhe: a natural language processing system for extracting cancer phenotypes from clinical records." *Cancer research* 77.21 (2017): e115-e118.