Cancer Registry Information Extraction via Transfer Learning

You-Chen Zhang[1], Ti-Hao Wang[2], Yi-Hsin Yang[3], Yan-Jie Lin[4], Chung-Yang Wu[1], Yu-Cheng Chang[1], Pin-Jou Lu[1], Chih-Jen Huang[5], Yu-Tsang Wang[6], Sheau-Fang Yang[7] Kuan-Chung Hsiao[3], Ko-Jiunn Liu[3], Li-Tzong Chen[3], Tsang-Wu Liu[3*] I-Shou Chang[3*], Kun-San Clifford Chao[8*], Hong-Jie Dai[1,3,9*]

[1]Intelligent System Lab, College of Electrical Engineering and Computer Science, Department of Electrical Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan R.O.C.
[2]Department of Radiation Oncology, China Medical University Hospital, China Medical University, Taichung, Taiwan, R.O.C.
[3]National Institute of Cancer Research, National Health Research Institutes, Tainan, Taiwan, R.O.C.
[4]Institute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan R.O.C.
[5]Cancer Center, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Taiwan R.O.C.
[6]Division of Medical Statistics and Bioinformatics, Department of Medical Research, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Taiwan R.O.C.
[7]Department of Pathology, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Taiwan R.O.C.
[7]Department of Public Health, Kaohsiung Medical University, Taiwan R.O.C.
[8] Cancer Center, China Medical University Hospital, China Medical University, Taichung, Taiwan, R.O.C.
[9]School of Post-Baccalaureate Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan R.O.C.

Abstract

A cancer registry is a critical and massive database for which various types of domain knowledge are needed and whose maintenance requires labor-intensive data curation. In order to facilitate the curation process for building a high-quality and integrated cancer registry database, we compiled a cross-hospital corpus and applied neural network methods to develop a natural language processing system for extracting cancer registry variables buried in unstructured pathology reports. The performance of the developed networks was compared with various baselines using standard micro-precision, recall and F-measure. Furthermore, we conducted experiments to study the feasibility of applying transfer learning to rapidly develop a well-performing system for processing reports from different sources that might be presented in different writing styles and formats. The results demonstrate that the transfer learning method enables us to develop a satisfactory system for a new hospital with only a few annotations and suggest more opportunities to reduce the burden of cancer registry curation.

1 Introduction

Cancer is a main cause of mortality worldwide and has been the leading cause of death over several decades in our country. A cancer registry system has been established by Taiwan Society of Cancer Registry and supported by Ministry of Health and Welfare (MOHW) over 40 years. How to extract massive data concisely and maintain high quality

* Corresponding authors

Proceedings of the 3rd Clinical Natural Language Processing Workshop, pages 201–208
November 19, 2020.

continuously are critical issues and burdens of healthcare system. However, the maintenance of an individual cancer registry from patient healthcare trajectories needs different types of domain knowledge which is pronouncedly both labor-intensive and time-consuming. In addition, how to validate and integrate between different hospitals or between local healthcare resource and national database are crucial topics.

To facilitate the integration of models for a specific cancer, applying information technology tools to improve acquisition and classification of patients' healthcare trajectories can enable more accurate phenotyping of cancer information. Nevertheless, addressing the issues needs more cooperation both on information technology and medical expertise. In order to assist integration among the institutes, a national project was established under the Cancer Center Support Grant Program (CCSG) supported by MOHW. As the coordinator of this project, we conducted research studies and cooperated with several hospitals to establish a platform to work out a model system based on existing cancer data.

One major goal of this project is to apply natural language processing (NLP) techniques to automatically analyze unstructured data including surgical reports, pathology reports, oncology clinical notes, and laboratory findings that may not be easy to acquire or share across hospitals for specific cancers. Pathology reports are usually abundant and contain operative findings, general tumor information, pathological assessment, cancer staging, and end-results which need to be extracted and classified clearly. In the pilot study, we focus on tasks including the collection and de-identification of pathology reports, data annotation for developing and evaluating deep learning-based NLP systems to extract cancer registry variables from different hospital sites.

To standardize the annotation of pathology material for developing our NLP system, the variables and their definitions were defined by the consensus from expertise committee composed of hospital investigators and annotators. Furthermore, we applied transfer learning and conducted experiments to examine the performance of the developed neural networks on the cross-hospital pathology materials to gain insights on how effective and concise transfer learning can be. The results not only enable us to understand which layers of the developed network convey the most important parameters for transfer but also let us know how many annotations are needed for training a system for a new hospital to achieve reasonable performance.

Table 1: Datasets collected from two medical centers for this study.

Source	CMUH	KMUH
# of Reports	393	1615
Training Set	293	1515
Test set	100	100
Period	2007~2013	2009~2015

2 Method

2.1 Datasets

In the presented study, we primarily focused on the colorectal cancer, which is the third leading cause of cancer-specific death in Taiwan. We cooperated with two medical centers, namely China Medical University Hospital (CMUH) and Kaohsiung Medical University Chung-Ho Memorial Hospital (KMUH), to collect colorectal pathology reports and the data were excluded non-tumor reports as well as the reports without cancer registration data for compiling our corpora. Table 1 shows the grouping and the number of the collected datasets.

2.2 Corpus Construction

In order to produce high quality annotations for developing our system, we established a NLP working group focusing on the construction of high-quality corpora. For our purpose, the annotation process was conducted by eight annotators based on an annotation guideline developed by consulting the committee composed of hospital investigators and cancer registrars. According to the standard of American Joint Committee on Cancer, nine cancer registry variables were defined for extraction in order to achieving a better understanding and unified effects on pathological materials. Table 2 summarizes the nine variables defined for the colorectal cancer including stage classification (SC), pathological TNM classifications (TNM), the number of examined nodes (NE) and positive nodes (PN), tumor size (TS), histology types (H), and grades (G). The entire annotation process is elaborated as follows.

A preliminary consistency test was conducted by asking the annotators to individually annotate

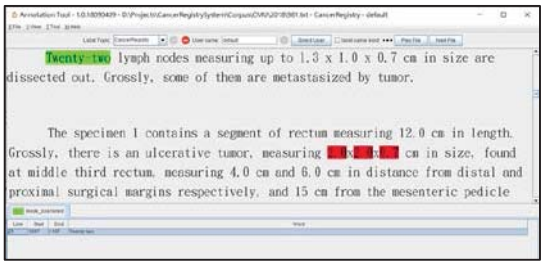

Figure 1: An example pathology report and the annotation tool used for annotation.

Afterwards a labeling meeting was organized to discuss issues and concerns encountered during the annotation process and the annotation guideline was adjusted according to the conclusion of the meeting. The above process was conducted iterative until they achieved an agreement above substantial. Finally, the remaining unlabeled datasets were evenly distributed to all annotators for labeling. The same annotation process was applied individually for the data collected from the two hospitals.

The aforementioned 100 annotation data generated by all annotators individually on the same reports were collected as the test set for evaluating the performance of the developed systems. They were combined by voting; only those annotations that were annotated by more than four annotators at the same time were kept. The other reports evenly annotated by annotators were collected as the training sets.

an identical set of 100 reports randomly selected from the collected datasets. All of them used the annotation tool (Figure 1) developed by our collaborator to conduct their annotations. We then measured their inter-annotation agreement by Kappa statistic (Viera & Garrett, 2005).

2.3 Cancer Registry Information Extraction with Different Approaches

For a given pathology report, our clinical toolkit (Dai, Syed-Abdul, Chen, & Wu, 2015) was employed to segment sentences and generate tokens based on MedPost (Smith, Rindflesch, & Wilbur, 2004). The numerical normalization method proposed by Tsai et al. (2006) was employed to reduce variations in numerical parts of each token. We then formulated the problem as a sequential labeling task and applied the IOB-2 tag scheme to encode the span information generated by annotators. All sequences including those that did not contain any annotations were included in the training set to train a neural sequence labeling network model whose architecture is briefly described as follows.

The input of the network is the pre-processed sequence of tokens in a pathology report and the output being the sequence of labels for each token. The input tokens was represented as a vector by concatenating the pre-trained word representations obtained by using GloVe (Pennington, Socher, & Manning, 2014) and RoBERTa (Liu et al., 2019). The parameters of the concatenated vectors were kept fixed during the training process.

Table 2: The nine cancer registry variables defined for this study.

Type	Description	Example
SC	Stage classifications including clinical, pathological, post-therapy/neoadjuvant therapy, retreatment/recurrence and autopsy	p., yp., rp., a., c.
T	Size or contiguous extension of the primary tumor	Primary tumor (T): Tx, T0, Tis, T1, T1, T2, T3, T4a, T4b
N	The absence, or presence and extent of cancer in the regional draining lymph nodes	Regional lymph nodes (N): Nx, N0, N1a, N1b, N1c, N2, N2a, N2b
M	The absence or presence of distant spread or metastases	Distant Metastasis (M): M0, M1, M1a, M1b
NE	Regional lymph nodes examined	Any numeric values
PN	Regional lymph nodes positive	Any numeric values
TS	Size of tumor	Any numeric values
H	Histology	Adenocarcinoma
G	Tumor grade; a measure of how abnormal the cancer cells look under the microscope.	Description likes: well differentiated, and undifferentiated

The concatenated representation was then feed to a fully connected layer (denoted as FC1) along with a variational dropout before passing the embeddings into the bidirectional long-short term memory (BiLSTM) network with one layer consisting of 256 hidden nodes. The output of the BiLSTM layer goes through another fully connected layer (denoted as FC2) to generate an output of a size equal to the number of the classes, which becomes the input of the inference layer in which a conditional random field (CRF) layer was used to model the dependencies between labels in neighborhoods with the Viterbi loss to jointly decode the best chain of labels for the given sequence.

In addition to the aforementioned architecture, we implemented the following baselines for comparison:

- Dictionary-based approach: For a given token, output the most frequent assigned tag estimated on the training sets.
- Support vector machine (SVM): Formulate the task as a token-based classification task and apply SVM with a polylinear kernel to learn a classification model.
- CRF: The normalized word features with a context window of three along with transition features were used for training a CRF model.
- BiLSTM: Similar to the aforementioned network architecture, but a linear layer was used instead of the CRF layer as the output layer.

All of the above neural networks were implemented by using PyTorch trained on NVidia Tesla P-100 GPUs. CRF was implemented by using CRF++[1] and scikit-learn[2] were used for the remaining implementations.

2.4 Transfer Learning for Extracting Information between Different Hospitals

Transfer learning (Pan & Yang, 2009) aims to learn a better model on a target domain by leveraging the knowledge previously learned from a source domain. In this study, the transductive transfer learning technology was applied by transferring the parameters in different layers of the BiLSTM-CRF model trained on the dataset of the source hospital to the target hospital by retraining the model with transferred parameters on the target hospital's dataset via fine-tuning. In our experiments, we didn't freeze any layers but fine-tuned all transferred parameters in different layers.

2.5 Experiment Configurations

We conducted three experiments to study the characteristics of the compiled corpora and the effectiveness of the developed models on the compiled corpora. The first compared the proposed model with the aforementioned baseline methods. The second examined the effectiveness of transfer learning and the last checked the robustness of the developed models under the evaluation of cross-corpus. The standard micro-precision (P), recall (R) and F-measure (F) were used to evaluate the models' outputs against the gold annotations.

For training the neural networks in all of our experiments, we randomly kept 50 reports in the training sets as the validation sets to determine the best performed models during the training process. The validation sets were not used in training. The mini-batch gradient descent along with the stochastic gradient descent algorithm (with a learning rate of 0.1, a momentum of 0.9 and a weight decay of 10^{-5}) was used for optimizing the parameters. Unless specifically described, the batch size and epoch were set to 2,048 and 150 respectively in the following experiments. The training process was early stopped if the learning rate was lower than 10^{-5}. For consistency, we used the same set of hyper-parameters and a fixed random seed across all experiments.

3 Results

3.1 Corpus Statistics

A total of 2,008 reports collected from the two hospitals were annotated. The final Kappa values estimated for CMUH and KMUH are 0.802 (substantial) and 0.914 (almost perfect) respectively. Table 3 shows the detail statistics of the compiled corpora. As one can see that the size of KMUH is much larger than that of CMUH. Although the size of the KMUH corpus is much larger than that of CMUH, the annotations for pathological M is much less in KMUH. It's because that pathological M stage need the other

[1] https://taku910.github.io/crfpp/

[2] https://scikit-learn.org/

Table 3: Corpus statistics for the compiled corpora used in this work.

	CMUH			KMUH		
Type	**Training**	**Test**	**Total**	**Training**	**Test**	**Total**
Histology	558	136	694	4,517	273	4,790
Grade	519	140	659	4,410	265	4,675
Numbers of examined nodes	623	189	812	2,021	153	2,174
Numbers of positive nodes	554	177	731	2,021	153	2,175
Staging classification	670	161	831	1,441	99	1,540
Pathological T	400	122	522	1,440	99	1,539
Pathological N	380	122	502	898	61	959
Pathological M	373	124	497	6	0	6
Tumor size	1,112	383	1,495	1,606	115	1,721
Numbers of reports	293	100	393	1,515	100	1,615
Numbers of sentences	21,229	5,699	26,928	63,887	3,966	67,853
Numbers of sentences with annotations	2,348	596	2,944	9,007	578	9,585
Numbers of annotations	5,189	1,554	6,743	18,360	1,218	19,578

reports (*e.g.*, image reports from other examination division) to conclude the outcome, the pathological M stage was shown inconclusive results on the current pathological data frequently in KMUH.

3.2 Performance Comparison with Different Methods

In the first experiment, we trained the developed models on the two training sets separately and evaluated their performance on the test sets of the two hospitals. The results were illustrated in Table 4.

In general, the developed models performed better on the KMUH test set which may be owing to the larger numbers of training samples. The CRF model achieved a comparable F-score on the KMUH test set but its F-score is lower than that of BiLSTM-CRF by 0.214 on the CMUH test set. Table 5 shows the detail results for the nine annotation types of the BiLSTM-CRF model on the two test sets. Overall, the developed networks demonstrated promising F-scores for all items.

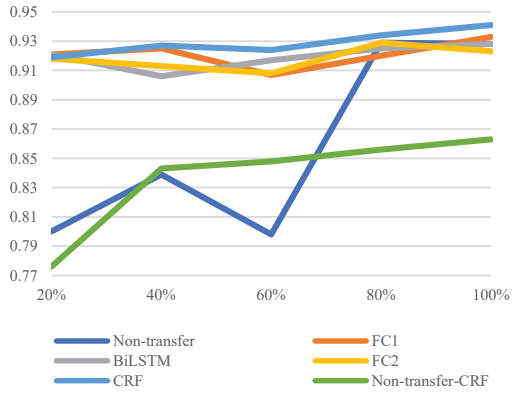

Figure 2: Impact of F-score by fine-tuning the models with the parameters up to each layer pre-trained on KMU on the varied sizes of the CMU training set.

3.3 The Effect of Transfer Learning

In this experiment, we would like to gain insights on what extent transfer learning improves the performance on the cross-hospital datasets. We used KMUH as the source dataset since its size is

Table 4: Performance comparison among different approaches.

	CMUH			KMUH		
Type	**P**	**R**	**F**	**P**	**R**	**F**
Dictionary-based	0.71	0.51	0.59	0.61	0.48	0.54
SVM	0.69	0.42	0.53	0.8	0.55	0.65
CRF	0.950	0.790	0.863	0.967	0.983	0.975
BiLSTM	0.823	0.638	0.719	0.975	0.975	0.975
BiLSTM-CRF	0.943	0.908	0.925	0.977	0.975	0.976

Table 5: Detail precision, recall and F-score for each cancer registry item of the BiLSTM-CRF model.

Type	CMUH		
	P	R	F
G	0.843	0.879	0.860
H	0.810	0.875	0.841
NE	0.973	0.968	0.971
PN	0.994	0.938	0.965
SC	0.946	0.988	0.967
TS	0.969	0.812	0.884
T	1.000	0.992	0.996
N	0.918	0.918	0.918
M	1.000	0.944	0.971
Type	**KMUH**		
G	0.996	0.996	0.996
H	0.968	0.985	0.976
NE	1.000	0.961	0.980
PN	0.981	1.000	0.990
SC	0.970	0.990	0.980
TS	0.991	0.913	0.950
T	0.921	0.939	0.930
N	0.952	0.967	0.959
M	n/a	n/a	n/a

larger than that of CMUH. We conducted experiments to examine the effect of transfer knowledge learned from KMUH to CMUH by 1) analyzing the importance of each layer of the developed neural networks, and 2) quantifying the performance gain by varying the sizes (20%~100%) of the CMUH training set when we fine-tuned the model pre-trained on KMUH. Note that because the size of the 20% CMUH dataset is quite small, we reduced the batch size to 512 for this case.

Figure 2 shows the results. Here "Non-transfer" refers to that we only used the reduced sizes of the CMUH training set to develop the BiLSTM-CRF models without relying on any pre-trained parameters. "FC1" initialized the learned parameters of the FC1 layer of the BiLSTM-CRF model by adopting the pre-trained parameters on the KMUH corpus, "BiLSTM" further included the learned parameters of the BiLSTM layer of the source model and so on. Consider the comparable results achieved by CRF models, we also include the configuration "Non-transfer-CRF" in which we trained several CRF models by using the corresponding reduced CMUH datasets.

In Figure 2, we can observe that with more numbers of the training samples used, the performance can be apparently improved for the 'Non-transfer' models. However, the improvement for the CRF models is relatively flat comparing with that of the neural networks. On the other hand, even with only 20% of the CMUH training set, the models learned with transferred parameters achieved satisfactory F-scores, which outperformed the 'Non-transfer' models trained on more training samples (being equal or less than 60%) of the full CMUH training set. The above results give us an insight that we can exploit the parameters of the neural networks learned from source hospitals to rapidly develop a reliable system relying on a small annotated dataset to boost the annotation process in the new hospital for creating and evaluating a customized system.

The results shown in Figure 2 also reveal the importance of parameters of each layer of the developed model in the manner of transfer learning. We can observe that transferring parameters of all layers in general leading to slightly better F-scores, but transferring the parameters of the first layer only is almost as efficient as transferring all. The result is consistent with the observations of other previous works (Giorgi & Bader, 2018, 2020; Lee, Dernoncourt, & Szolovits, 2018) and the hypothesis that the lower layers of a neural network learn generic features and the higher layers learn task-specific (or we can say that hospital-specific) features.

3.4 Cross-corpus Evaluations

To assess the performance of the developed model in a more realistic setup, we conducted cross-

Table 6: Cross-corpus evaluation among different approaches.

	CMUH			KMUH		
Method	**P**	**R**	**F**	**P**	**R**	**F**
CRF	0.737	0.242	0.364	0.663	0.314	0.426
BiLSTM-CRF	0.631	0.376	0.472	0.925	0.483	0.634
Transferred BiLSTM-CRF	0.944	0.938	0.941	0.932	0.644	0.762

datasets experiments. For this purpose, we used the dataset from one hospital for training, and the dataset from another for testing. The experiments provide an estimate of the cross-hospital generalization ability of the developed models.

Table 6 shows the results. Given that both corpora were annotated by the same annotators under the same annotation guideline, we can still see the generality of the developed models is not well; a larger drop in performance can be found on both datasets. The results exhibited that the format and the writing styles of the descriptive pathology in surgical biopsy reports across hospitals are heterogeneous in real-world scenarios.

We also estimated the performance of the transferred model on its source dataset in Table 6. The result illustrates an apparent drop of F-score from 0.976 to 0.762 on the KMUH test set. The results demonstrated that the developed systems suffered the catastrophic forgetting problem (French, 1999) which is now known to be a challenge for artificial neural networks when the network is trained sequentially on multiple tasks because the weights in the network that are important for the original task are now changed to meet the objectives of the new task.

4 Conclusions

In this work, we investigated the feasibility of applying transfer learning via neural networks on the task of extraction cancer registry information from cross-hospital pathology reports. Because the writing styles and formats of the pathology reports is different in each hospital, to estimate the requirements of the number of annotated datasets when we migrate from one hospital to the others and iteratively improve the effectiveness of the developed systems, we conducted experiments to quantify the impact of transfer learning on the datasets collected from two hospitals. From the evaluations of the results, we confirmed that when transfer learning is adopted, the model pre-trained on a source hospital can be trained with fewer annotations of the target hospital and achieve satisfactory performance as when the full training set of the target hospital is used. The results suggest us to apply the transfer learning techniques for developing a customized system for a new hospital with only a few annotations. We will develop method to estimate the required numbers of annotations based on the language properties of the narrative reports and the characteristics of the developed neural networks. Furthermore, our experiment results also reveal challenges requiring to be addressed including the generalizability and catastrophic forgetting problem, which should be addressed in the future.

Acknowledgement

This study was supported by the Ministry of Health and Welfare [grand number: MOHW109-TDU-B-212-134026] and the Ministry of Science and Technology of Taiwan [Grant numbers: MOST 109-2221-E-992-074-MY3].

References

Dai, H.-J., Syed-Abdul, S., Chen, C.-W., & Wu, C.-C. (2015). Recognition and Evaluation of Clinical Section Headings in Clinical Documents Using Token-Based Formulation with Conditional Random Fields. *BioMed Research International, 2015*.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences, 3*(4), 128-135.

Giorgi, J. M., & Bader, G. D. (2018). Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics, 34*(23), 4087-4094.

Giorgi, J. M., & Bader, G. D. (2020). Towards reliable named entity recognition in the biomedical domain. *Bioinformatics, 36*(1), 280-286.

Lee, J. Y., Dernoncourt, F., & Szolovits, P. (2018). *Transfer Learning for Named-Entity Recognition with Neural Networks.* Paper presented at the Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering, 22*(10), 1345-1359.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014), 12*, 1532-1543.

Smith, L., Rindflesch, T., & Wilbur, W. J. (2004). MedPost: A Part of Speech Tagger for BioMedical Text. *Bioinformatics, 20*(14), 2320-2321. doi:10.1093/bioinformatics/bth227

Tsai, R. T.-H., Sung, C.-L., Dai, H.-J., Hung, H.-C., Sung, T.-Y., & Hsu, W.-L. (2006). NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. *BMC Bioinformatics, 7*(Suppl 5), S11.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med, 37*(5), 360-363.