Automated Histologic Grading from Free-Text Pathology Reports Using Graph-of-Words Features and Machine Learning

Hong-Jun Yoon, Larry Roberts and Georgia Tourassi, *Member, IEEE*

***Abstract*— Traditional *n-gram* feature representation of free-text documents often fails to capture word ordering and semantics, thus compromising text comprehension. Graph-of-words, a new text representation approach based on graph analytics, is a superior method overcoming the limitations by modeling word co-occurrence. In this study, we present a novel application of graph-of-words text description for automated extraction of histologic grade from unstructured pathology reports. Using 10-fold cross-validation tests, the proposed approach resulted in substantially higher macro and micro-F1 scores with undirected graph-of-words features, compared to traditional *bi-gram* text features. Our feasibility study demonstrated that graph-of-words is a highly efficient method of text comprehension for information extraction from free-text clinical documents.**

I. Introduction

Automatic information extraction from natural language text is one of the essential tasks of big data science, attracting much attention from research communities dealing with massive volumes of complex unstructured text data [1]. This is particularly true for clinical applications where manual annotation by trained personnel is necessary but highly costly and often error prone [2]. Cancer surveillance is one of these fields for which artificial intelligence-based information extraction is highly desirable since fast and accurate text data comprehension is crucial for monitoring cancer occurrence and progression by histologic type across the population [3].

The Bag-of-Words (BOW) model [4] is one of the most popular word representation methods for text categorization. The idea is to quantize each word or combination of words, referred as *n-gram* and represent the text document by the number of the *n-gram* occurrences. The occurrence of each *n-gram* is then used as a feature for document classification. The BOW model is widely applied for information extraction and classification of electronic health records and cancer pathology reports. Ogren et al. introduced ClearTK that composes BOW model where the *n-gram* features were extracted corresponding to the biomedical part-of-speech tagging [5]. Koopman et al. utilized terms and *n-grams* to identify cancer-related causes of death from death certificates [6]. Martinez and Li evaluated BOW model alone as well as in conjunction with other information extraction models such as the Unified Medical Language System (UMLS) lexical database to extract information from pathology reports [7]. Kavuluru applied unigrams, bigrams, and named entities as features to identify primary sites from cancer pathology reports [8].

A key limitation of the BOW model is its inability to effectively handle ambiguity of expression variations, such as "word inversion" (e.g. "RIGHT BREAST CORE BIOPSY" and "BREAST, RIGHT, CORE BIOPSY") and "subset matching" (e.g., "RIGHT LOBE" and "RIGHT UPPER LOBE"). For example, the phrases "BREAST (LEFT 2:00 BX)" and "LEFT BREAST MASS AT 2:00" are referring to the same primary site of the breast cancer, ICD-O-3 code C50.4, but may not be recognized as the same category by a classifier trained using conventional *n-gram* features. One may argue that the set of three independent unigrams {LEFT, BREAST, 2:00} could describe the document better instead of a single trigram. While such simplified representation may seem reasonable, it can easily lead to document misclassification due to the inherent complexity of the cancer pathology reports where such unigrams may exist elsewhere within the report but not in reference to the primary cancer. Graph-of-Words (GOW) [9] is a method introduced to overcome such issues because graphs are useful to represent the interaction between entities of words.

In this study, we employed a Bag-of-Graphs (BOG) model which counts the number of occurrence of GOWs in a document to identify histologic grades from cancer pathology reports. To the best of our knowledge, this approach has not been applied before in the context of this application. The article is organized as follows. First, we review the study dataset and the BOG model in Section 2. Classification performance of the BOG model for automated extraction of histologic grade information from the pathology report dataset is presented in Section 3 along with comparative results to conventional methods applied before for the same task. Lastly, we discuss the study findings and future direction in Section 4.

II. Methods

A. Cancer Pathology Report Data

Pathology reports are unstructured text documents containing information about human tissue specimens. They are a standard component of clinical reporting and management of cancer patients. In addition, cancer pathology reports are a primary source of information for the Surveillance, Epidemiology, and End Results (SEER)

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of the manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

H. Yoon, L. Roberts, and G. Tourassi are with the Health Data Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA (email: yoonh@ornl.gov, robertslw@ornl.gov, tourassig@ornl.gov).

program [10]. The SEER program is the premier population cancer surveillance program covering 30% of the US population. It is an important national resource for monitoring cancer outcomes across demographic groups, geographic regions, and time. Furthermore, SEER provides unique insights into the impact of oncology practice outside the clinical trial setting. The information collected includes data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, the first course of treatment, and follow-up for vital status.

B. Cancer Histologic Grades

Cancer histologic grading is based on the microscopic appearance of a malignant neoplasm [11]. In general, higher grade means that there is lesser degree of differentiation and the worse the biologic behavior of a malignant neoplasm will be. A well- differentiated neoplasm is composed of cells that closely resemble the cell of origin, while a poorly differentiated neoplasm has cells that are difficult to recognize as to their cell of origin. Most histologic grading systems include three or four grades. The grade of cancer depends on what the cells look like under a microscope. In general, a lower grade indicates slower-growing cancer and a higher grade indicates a faster-growing one. Grade 1 cancer cells resemble normal cells and are not growing rapidly, whereas grade 3 cancer cells look abnormal and may grow and spread aggressively.

This study was based on 661 pathology reports annotated by expert cancer registrars. The dataset included breast and lung cancer pathology reports. Definitions of cancer histologic grades and the number of cases employed for this dataset are listed in Table I.

TABLE I. DEFINITIONS OF CANCER HISTOLOGY GRADES AND NUMBER OF CASES EMPLOYED IN THE STUDY

Grade	Definition	# cases
I	Well differentiated	129
II	Moderately differentiated	240
III	Poorly differentiated	275
IV	Undifferentiated / Anaplastic	17

C. Graph-of-Words

We adopted *Graph-of-Words* (GOW) as a representation of documents of natural language text. The graph nodes are unique words in the documents and edges represent co-occurrences between the words within distance *d*. An example of text parsing and the GOW representation of a sample paragraph in a pathology report is illustrated in Fig. 1.

Intuitively, graph representation provides flexibility and robustness on representing natural language text, compared to the traditional *n-gram* approach. For example, with respect to tumor histologic grade statements, one may observe linguistic variability such as

"histologic grade: 3",

"histologic grade (mbr): 3", and

"histologic grade (g1-3): 3".

Note that a single *bi-gram* "grade 3" is not feasible to represent all those variations, but is adequate with a graph ("grade" – "3", $d > 1$).

Note also that the "directed" GOW's have edges that count "word order", i.e., "grade" – "3" is not the same as "3" – "grade". On the other hand, the "undirected" GOW's uses pairs of edges in both directions, thus we can expect flexibility with "word inversion" cases.

Pre-processing was applied as follows: contents were lower-cased, characters except for alphanumeric codes and whitespaces were removed, and stopwords [12] were deleted. We also applied some heuristic rules along with the pre-processing steps. First, the words "i", "ii", "iii", and "iv" in lines containing the words "grade" or "stage" were converted into "1", "2", "3", and "4". Since pathology reports are written by the third-person point-of-view we can assume that "i" is not used by the first-person singular in the context of the pathology reports. Second, words "g1", "g2", "g3", and "g4" were converted into "1", "2", "3", and "4 because we observed that some reports used "g1", "g2", "g3", and "g4" as a grading convention. Lastly, the stopwords "no" and "not" were not removed, which may inverse the meaning of sentences, e.g., "no metastatic carcinoma seen" and "not well differentiated".

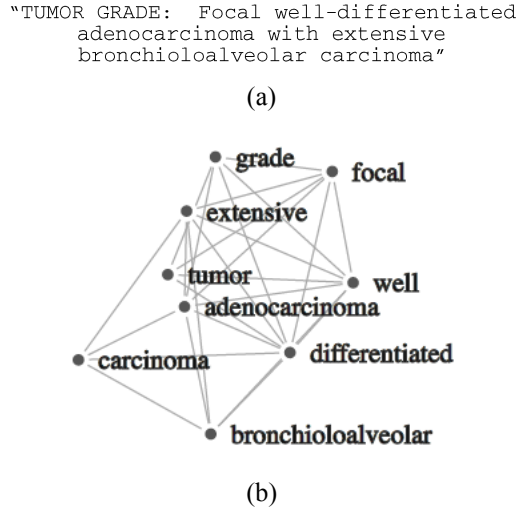

Figure 1. Example of undirected graph-of-words representation of a (a) sample text from a pathology report and (b) its node and edge with maximum word distance d=4.

D. Bag-of-Graphs Model

Document classification requires vectorization that converts documents of natural language texts into numerical feature vectors. In this study, we applied a Bag-of-Graphs (BOG) model, which first learns GOW's from all documents, and then it models each document by counting the number of occurrences that each GOW appears. Those numbers allow us to compare documents and gauge their similarities.

We selected GOW's that are best to describe the texts of pathology reports based on their document frequency in each histologic grading class, followed by their odds ratio between the documents of each histologic grading class and the rest pathology report corpus.

E. Performance Measure

We applied F-scores for evaluating classification performance, which is widely used for information retrieval classification tasks. Let $TP_i = M_{ii}$, $FN_i = \sum_{j \neq i} M_{ji}$, and $FP_i = \sum_{j \neq i} M_{ij}$, represent true positive, false negative, and false positive decisions respectively where M_{ij} implies the number of pathology reports from class i assigned to class j. Then one can compute,

$$Precision_i = \frac{TP_i}{TP_i + FP_i}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i}$$

The F-score of class i is the harmonic mean of the precision and recall of the class,

$$F_\beta\text{-score}_i = (1 + \beta^2) \cdot \frac{Precision_i \cdot Recall_i}{\beta^2 \cdot Precision_i + Recall_i}$$

where β is a weight factor assigning higher importance to recall if $\beta > 1$ and to precision otherwise. Typically, $\beta = 1$, called balanced F-score, or F_1-score. Macro-average F_1-score is defined as

$$Macro\text{-}F_1 = \frac{1}{N} \cdot \sum_i F_1\text{-}score_i$$

where N is the number of classes. Micro-average F_1-scores are defined as follows

$$Precision' = \frac{\sum_{i=1}^{N} TP_{class_i}}{\sum_{i=1}^{N} \left(TP_{class_i} + FP_{class_i}\right)},$$

$$Recall' = \frac{\sum_{i=1}^{N} TP_{class_i}}{\sum_{i=1}^{N} \left(TP_{class_i} + FN_{class_i}\right)},$$

$$Micro\text{-}F = \frac{2 \cdot Precision' \cdot Recall'}{Precision' + Recall'}.$$

The macro-average method measures how the classifier performs overall across the dataset. On the other hand, the micro-average is a meaningful measure when there is imbalanced representation of the classes in the dataset.

III. Experimental Results

We performed a 10-fold cross-validation experiments based on the 661 pathology reports per each word distance *d*, and each directed and undirected GOW's. Note that the directed GOW of *d*=1 is equivalent to the traditional *bi-gram* representation.

Table II lists both macro and micro F_1 scores of the classification results and optimal parameters found. We compared three conventional classifiers typically used for natural language processing applications. The implementation was on Python 2.7.12 environment, numpy 1.11.1, and scikit-learn 0.18 libraries.

TABLE II. Macro and Micro F_1-scores of Histology Grade Classification on the Cancer Pathology Reports for three classification schemes, Best performance was achieved with the Random Forest Classifier, Undirected Graph, and the Maximim Word Distance 4

	Naïve Bayes		Logistic Regression		Random Forest	
d	**Macro F_1**	**Micro F_1**	**Macro F_1**	**Micro F_1**	**Macro F_1**	**Micro F_1**
	Directed					
1	0.455	0.523	0.433	0.531	0.404	0.520
2	0.562	0.661	0.586	0.669	0.588	0.682
3	0.585	0.644	0.652	0.743	0.663	0.735
4	0.486	0.543	0.650	0.737	0.660	0.737
5	0.509	0.554	0.617	0.735	0.660	0.720
6	0.660	0.703	0.632	0.723	0.653	0.723
	Undirected					
1	0.739	0.806	0.804	0.867	0.808	0.867
2	0.668	0.708	0.768	0.841	0.797	0.846
3	0.689	0.737	0.779	0.844	0.827	0.864
4	0.674	0.707	0.774	0.832	***0.833***	***0.884***
5	0.668	0.705	0.775	0.841	0.813	0.865
6	0.674	0.708	0.794	0.818	0.820	0.840

Table II demonstrates that the performance of GOW representation is far superior to the traditional *bi-gram* feature (directed GOW, *d*=1), regardless of the choice of classifiers. F_1 scores based on undirected graphs performed consistently better than that of directed ones, suggesting that the flexibility on "word order" is more effective for the classification task at hand. Longer search distance of words *d* gave a better chance to capture the variation of word expressions, but further increase reduced classification performance suggesting that large distances *d* may introduce ambiguity. We found that the optimal classification performance was achieved when *d* is around 3 and 4.

We looked closely into the details of the classification results by deriving the confusion matrix, listed in Table III, as well as the selected GOW's, listed in Table IV. We achieved the best results achieved using the Random Forest classifier with *d*=4. Table IV illustrates that the suggested BOG feature selection is highly efficient, since the algorithm captured important keywords that define the characteristics of histologic grades, such as the expressions of cell differentiation and cancer histologic grades. Those keywords possessed relatively high odds ratio.

TABLE III. Confusion Matrix of the Cancer Histology Grade Classification Based on Bag-of-Graphs Feature Representation and Random Forest Classifier, Predicted Labels in Columns and Actual Labels in Rows, with Precision and Recall by Class of Histologic Grade

	Grade 1	**Grade 2**	**Grade 3**	**Grade 4**	**Recall**
Grade 1	110	16	3	0	0.853
Grade 2	18	207	15	0	0.863
Grade 3	2	15	258	0	0.938
Grade 4	0	2	6	9	0.529
Precision	0.846	0.863	0.915	1.000	

TABLE IV. CHOICES OF GRAPH-OF-WORDS THAT SPECIFY CANCER HISTOLOGIC GRADES OF PATHOLOGY REPORTS AND THEIR ODDS RATIO,.

Word 1	Word 2	Odds Ratio
carcinoma	undifferentiated	1178.83
biopsy	undifferentiated	351.27
poorly	differentiated	64.77
grade	1	16.87
well	differentiated	13.98
moderately	differentiated	12.05
grade	2	6.19
grade	low	5.89
grade	high	5.51
grade	intermediate	4.23

IV. DISCUSSION

In this paper, we applied GOW feature representation to the classification of histologic grades on cancer pathology reports. These are highly complex free-text documents with substantial linguistic variability even when using similar medical terminology. Feasibility was evaluated by a comparative study using 10-fold cross-validation experimental design comparing traditional *bi-gram* and the proposed GOW feature, both directed and undirected graphs, using three different classification schemes; naïve Bayes, logistic regression, and random forests. Both macro-F_1 and micro-F_1 scores were calculated as performance measures of the classifiers.

The results clearly demonstrated that the proposed GOW features performed substantially better than the traditional *bi-gram* features, implying that the GOW is an effective means of description for free-form natural language pathology reports. Furthermore, we observed that undirected graphs lead to superior classification performance than directed graphs, suggesting that the undirected graph is more effective for describing word order variabilities. The search distance of words is also an effective way of describing variability of natural language text expressions.

In conclusion, the promising results encourage further extension of the approach for other important information extraction tasks such as primary cancer site category classification. Additionally, we will exploit graph analytics such as graph similarity to achieve more effective and robust features for text comprehension.

ACKNOWLEDGMENT

This work has been supported in part by the Joint Design of Advanced Computing Solutions (JDASC4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health.

The authors wish to thank Valentina Petkov of the Surveillance Research Program from the National Cancer Institute and the SEER registries at HI, KY, CT, NM and Seattle for the de-identified pathology reports used in this investigation.

REFERENCES

[1] S. Soderland, "Learning information extraction rules for semi-structured and free text," Machine learning, vol. 34, no. 1-3 pp. 233-272. 1999.

[2] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research," Yearb Med Inform, vol. 35 pp. 128-44, 2008.

[3] I. Spasić, J. Livsey, J. A. Keane, and G. Nanadić, "Text mining of cancer-related information: review of current status and future directions," International journal of medical informatics, vol. 83, no. 9 pp. 605-623, 2014.

[4] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," In European conference on machine learning, pp. 137-142. Springer Berlin Heidelberg, 1998.

[5] P. V. Ogren, P. G. Wetzler, and S. J. Bethard, "ClearTK: A UIMA toolkit for statistical natural language processing," Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP, pp. 32-38, 2008.

[6] B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, and N. Grayson, "Automatic ICD-10 classification of cancers from free-text death certificates," International journal of medical informatics, vol. 84, no. 11, pp. 956-965, 2015.

[7] D. Martinez and Y. Li. "Information extraction from pathology reports in a hospital setting," In Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1877-1882, ACM, 2011.

[8] R. Kavuluru, I. Hands, E. Durbin, and L. Witt, "Automatic extraction of ICD-O-3 primary sites from cancer pathology reports," In Clinical Research Informatics AMIA symposium, March 2013.

[9] F. Rousseau and M. Vazirgiannis, "Graph-of-word and TW-IDF: new approach to ad hoc IR," In Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pp. 59-68, ACM, 2013.

[10] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2013), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2016.

[11] M. T. Carriaga and D. E. Henson. "The histologic grading of cancer," Cancer vol. 75, no. S1, pp. 406-421, 1995.

[12] S. Bird, E. Klein, and E. Loper, Natural language processing with Python. O'Reilly Media, Inc., 2009.