POE: A Pathology Extraction Tool for Finding Attribute-Value Pairs in Glioma Pathology Reports.

Veronica E. Lynn, BA[1], Niranjan Balasubramanian, PhD[1], Tahsin Kurc, PhD[1],
Joel Saltz, MD, PhD[1], Rebecca Jacobson, MD, MS[2],
[1]Stony Brook University, Stony Brook, NY; [2]University of Pittsburgh, Pittsburgh, PA

Introduction

Histopathologic features, identified by direct examination of neoplasms, play a critical role in cancer diagnosis. Diagnostic classifications, such as the WHO classification for glial neoplasms, provide guidance on the sub-classification and grading of cancer. For example, a high degree of cellularity and necrosis is associated with higher grade and worse outcomes in glial tumors. New integrative approaches that combine information from histologic, imaging and genomic features have the potential to advance methodologies for cancer classification[1] and will likely necessitate development of new classification schemes. These approaches will require scalable extraction of semantically rich features, potentially in very large datasets.

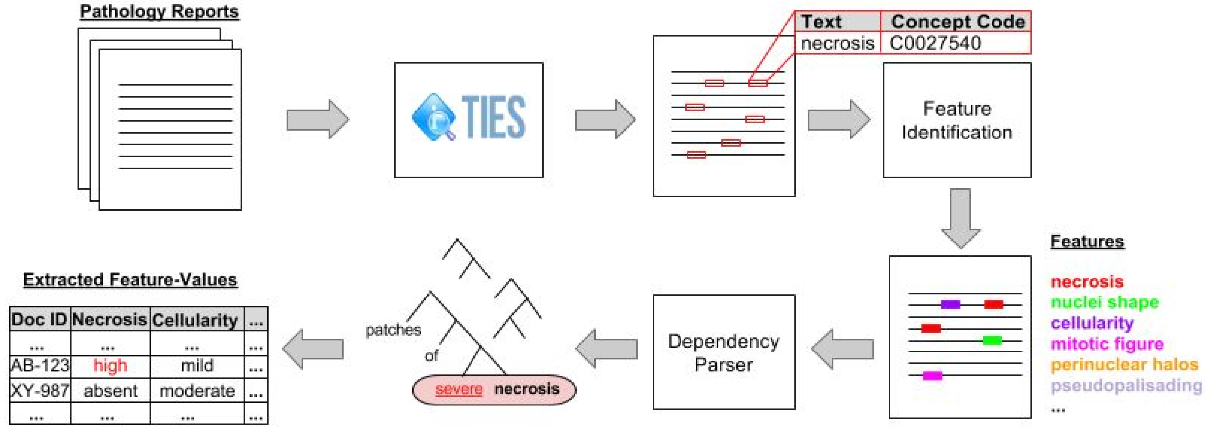
At present descriptions of histopathologic features are available in free text surgical pathology reports, and natural language processing is needed to extract them. In this case study, we develop and evaluate a pathology extraction tool, ***POE***, which builds on the TIES framework[2], for Glioma pathology reports. Many existing NLP tools [2,3,4] identify medically relevant entities drawn from medical terminologies; TIES in particular provides a carefully hand-crafted set of cancer-specific entities. Tools such as TIES provide additional markup including negative observations and findings which have diagnostic value. In addition to these valuable annotations, we also need entity-attribute relationships to capture salient morphologic features, and the grading of observations, which has critical diagnostic value. For example, we may want to know whether *cellularity* was high or low to assess whether the specimen is a low or high-grade glioma. While it is relatively easy to find modifiers, determining which entity they modify is difficult. The terse nature of the reports means there are often many possible candidates, making this a hard task. *To address these challenges, POE uses i) a feature classifier to identify relevant histopathologic features, and ii) a feature grade extractor that uses a syntactic dependency parser to locate the modifiers for each feature mention and map it to a grade value.*

Method

On its own, TIES processes documents to identify entities and annotates their text spans with their concept name, concept code, and negation information. POE extends this by using the entity annotations to identify specific morphological features and link them with modifiers, which are then mapped to a grade value. This process is explained in more detail in Figure 1. In total, POE can extract 12 features, chosen due to being commonly considered when making a diagnosis, which can be assigned values based on severity (*absent, present, mild, mild to moderate, moderate, moderate to high,* or *high*) or shape (*round, round to oval*, or *oval*). Some examples are given in Figure 1.

Results

We evaluated POE using 473 pathology reports (250 glioblastoma, 223 low-grade glioma) collected from the Cancer Genome Atlas (TCGA). For these reports, the final diagnosis is given but the feature values must be extracted from the text. We therefore manually annotated 213 of these reports (132 glioblastoma, 81 low-grade glioma) to obtain the feature values.



Text	Concept Code
necrosis	C0027540

Doc ID	Necrosis	Cellularity	...
...	...	...	...
AB-123	high	mild	...
XY-987	absent	moderate	...
...	...	...	...

Figure 1. POE pipeline: (1) Reports are given to TIES to extract annotated text spans. Each annotation contains concept information including a concept name, concept code, and negation. (2) Annotations are classified to determine what feature, if any, they represent. The text spans and their concept codes are provided as features to the classifier. (3) Sentences containing features are dependency parsed. The parse tree is searched to identify modifying words or phrases in the sibling and descendant nodes. (d) All modifiers linked to a particular feature are resolved into a single, final value and output into a table.

POE's feature identification classifier was evaluated through cross validation on the manually annotated set. It was able to determine if a feature was mentioned within a report with an F1 of **87.2**. We also evaluated POE's extractor for its ability to assign values for each feature. Across all possible values, POE obtained an accuracy of **52.2%** at assigning the correct grade, a nine percent improvement (statistically significant) over a nearest-modifier heuristic. The drop in accuracy reflects the difficulty of assigning a grade compared with identifying feature mentions. Grade assignment typically requires a deeper semantic understanding of the text, necessitating a more sophisticated approach.

We demonstrate the utility of POE's extracted features by using them to train a random-forest classifier to predict glioma type. The classifier is evaluated through cross validation using the full 473 reports, all of which have been automatically annotated using POE. We achieve an accuracy of **76.3%** (80.4% glioblastoma, 71.7% low-grade glioma), compared to a random chance baseline of 52.9%.

Discussion

Our preliminary evaluations show that POE's extensions to TIES can effectively extract relevant histopathologic features from glioma pathology reports. The evaluations also show that the extracted histopathologic features have some diagnostic value, motivating investigations into their utility for assessing new classification schemes. Effective extraction of histopathologic features from free-text reports is a precursor to determining how these features relate to each other, to image-based features and with outcomes. It will be a critical component in developing robust, integrative approaches for evaluating new classification schemes and designing effective therapies. Our next steps include (i) addressing extraction issues which stem from stylistic variations and the terse nature of the reports, and (ii) scaling the feature identification for other types of cancers. To promote further investigations in the community, we have released a dockerized version of POE (https://hub.docker.com/r/sbubmi/pathomics_morphex) that can produce a list of histopathologic feature-value pairs from pathology reports.

References

1. Colen, Rivka, et al. NCI Workshop Report: Clinical and Computational Requirements for Correlating Imaging Phenotypes with Genomics Signatures. Translational Oncology, 7(5): 556-589, Oct 2014.
2. Crowley RS, Castine M, Mitchell KJ, Chavan G, McSherry T, Feldman M. caTIES - A Grid Based System for Coding and Retrieval of Surgical Pathology Reports and Tissue Specimens In Support Of Translational Research. J Am Med Inform Assoc. 2010 May 1;17(3):253-64. PMCID: PMC2995710
3. 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. Journal of the American Medical Informatics Association. 2010.
4. Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium 2001 (p. 17). American Medical Informatics Association.