A Knowledge-based Approach to Information Extraction from Surgical Pathology Reports

Kevin J. Mitchell, MS[1], Rebecca S. Crowley, MD, MS[2], Dilip Gupta, MD[2], and John Gilbertson, MD[2]
[1] Benedum Oncology Informatics Center, University of Pittsburgh Medical Center
[2]Center for Pathology Informatics, University of Pittsburgh, Pittsburgh, PA

ABSTRACT

We describe the development of a prototype system for knowledge-based information extraction from surgical pathology reports. The current system includes abstract problem solving methods and a frame-based knowledge representation of body parts, procedures, diseases, and findings for prostate and breast cases. The system currently extracts the organ, procedure, and diagnoses, and sets an agenda of goals for further processing. A potential advantage of this approach is the ability to increase specificity of information extraction.

BACKGROUND

Surgical Pathology reports represent an enormously valuable source of information for clinical and basic science researchers. Unfortunately, almost all of the important data in these reports is encoded as structured, free text. Surgical Pathology reports indicate the organ, site, procedure, histologic diagnoses and supporting findings, presence or absence of important prognostic features, and results of ancillary studies such as molecular markers and immunohistochemistry. Reports also act as tissue locators by indicating the presence of tissue blocks, frozen tissue and other resources, and by identifying the relationship of the tissue block to significant landmarks (such as tumor margins).

The Shared Pathology Information Network (SPIN) – a consortium project of the National Cancer Institute will establish the feasibility of a distributed network of institutions permitting access to information in the free text of locally maintained, de-identified surgical pathology reports. The complete SPIN Project description is available at http://www.cancerdiagnosis.nci.nih.gov/spin/. SPIN will provide a powerful research resource to support hypothesis driven research and data-mining. In previous work[1] we have modified existing software to de-identify surgical pathology reports, and shown that it performs extremely well. The goal of the current work is to develop methods for extracting information from surgical pathology reports.

APPROACH

The linguistic complexity of narrative in this domain creates significant challenges to information extraction. Reports typically include frequent abbreviations, negation, synonymy, homonymy, polysemy, and context-sensitivity. Concepts are often expressed in such a way that use of regular expressions may not permit extraction with high specificity. For example, the value of the size of a tumor (a real number and a corresponding measurement unit) is almost always present without the name of the attribute ("size"). Often, other real numbers are present in the narrative, but have nothing to do with tumor size. We reasoned that contextual knowledge of the procedure, organ and diagnosis could be used to guide information extraction. For example, attempts to extract "size of tumor" could be limited to cases in which this measurement is applicable – i.e. cases of cancer resections. In these cases, the prior probability of finding the feature is already high.

Analysis of existing reports indicates that in almost all cases the necessary information for inferring body part and procedure resides on the first line of the Final Diagnosis section of the report. However syntactic matching alone is too rigid for the inherent variability in reporting practices. For example the concept "prostatectomy" often must be inferred from a combination of concepts (e.g. "prostate, resection"). We implemented a prototype knowledge-based information extraction system that uses abstract problem solving methods and a frame system to determine the most specific organ-specified procedure and diagnosis (e.g. benign biopsy of prostate), and then creates an agenda of extraction goals, specific to this information.

SYSTEM DESCRIPTION

The knowledge base consists of a set of frames derived from UMLS concepts using the UMLS Tab plug-in in Protégé-2000. Only instances related to prostate and breast cases were created for this prototype. The representation is taxonomic, producing a hierarchy of diseases, body parts, procedures and report findings. Concepts and relationships not present in the UMLS were added. Expected findings are related to the organ, procedure, and diagnoses to which they apply. Procedures are related to disease categories for which they are indicted (e.g. radical prostatectomy is related to malignant neoplasms of prostate), and to the body part specimens that are generated. The findings (prognostic features) to be identified were modeled from existing descriptions of common data elements for prostate and breast pathology. The problem-solving methods are executed in Jess – a Java Production Rule System. Reports are sequentially fed into the system. A UMLS autocoder first extracts all UMLS concepts, and then matches them against the frame set to yield the most specified organ, procedure, and diagnosis possible, and to generate the agenda of extraction goals. Performance across one thousand prostate cases shows that extraneous concept codes are eliminated and agendas are appropriately set to guide template-based extraction. Future work will include recognition of negated concepts and methods to perform extraction based on the inferred agendas.

ACKNOWLEDGEMENTS

This work was supported by National Cancer Institute Grant UO1 CA91343 for the Shared Pathology Informatics Network (SPIN). We thank Jules J. Berman for use of his autocoder (http://65.222.228.150/jjb/parse.tar.gz)

REFERENCES

1. Dilip Gupta, Melissa Saul , John Gilbertson. Evaluation of a De-identification Software Engine: Progress Towards Sharing Clinical Documents and Pathology Reports. American Medical Informatics Association (AMIA), San Antonio November 12, 2002.