

Evaluation of Open Source Text Mining Tools for Cancer Surveillance:

Phase I: Understanding text mining and identifying tools

Prepared by: Wendy Scharber

NPCR-AERRO Technical Development Team

CDC/NCCDPHP/DCPC

2007

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

4/14/2009

Page 1 of 28

Evaluation of Text Mining Tools

Table of Contents

Background.....	3
Purpose of Project.....	4
Contents of this Report:	5
Methodology.....	5
Understanding Text Mining	6
Information Extraction and Information Retrieval.....	6
Definitions and Uses	7
Lexical Level	7
Syntactic Level.....	9
Semantic Level.....	11
Challenges in Identifying Terms	13
Term Variation.....	13
Acronyms and Abbreviations and Ambiguity.....	14
The Text Mining Process.....	14
Results of Literature Search	16
Open Source Tools	17
CaFE: Registry Case Finding Engine.....	17
Extended MedLEE: Medical Language Extraction and Encoding System.....	18
HITex: Health Information Text Extraction.....	19
caTIES: cancer Text Information Extraction System.....	20
Proprietary Software.....	20
Discussion	21
Tool Evaluation.....	21
caFE.....	21
MedLEE	22
HITex.....	22
ca_TIES	24
Conclusions.....	25
Action Plan.....	27

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

Background

Data mining has been in use within the cancer surveillance community for several years. The first central cancer registry was established in 1935 in Connecticut¹ and since that time, given the comprehensive and rich population-based registry data, epidemiologists and researchers have contributed significantly to understanding the cancer burden. A review of the following websites provides examples of the use of cancer registry data in research:

- <http://www.cdc.gov/cancer/npcr/datarelease.htm>
- <http://www.seer.cancer.gov/publications/>
- http://www.naaccr.org/index.asp?Col_SectionKey=11&Col_ContentID=462

Data mining is performed on structured data, which has an enforced composition. Formal definitions and data types are required for collecting structured data, and it resides in a fixed place within a record, file or document.²

In order to provide epidemiologists and researchers structured cancer data for data mining, an enormous amount of *unstructured* data need to be evaluated, synthesized and condensed into a coded format. Unlike structured data, unstructured data has no conceptual definition and no data type definition – a word

¹ <http://vvy.dph.state.ct.us/OPPE/hptumor.htm>

² G Weglarz. Two Worlds of Data – Unstructured and Structured. DM Review Magazine, September 2004.

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

is simply a word.³ Converting unstructured data to structured data is performed manually by cancer registrars who read all of the pertinent information to identify diagnoses of cancer and abstract data into a structured format using established standards⁴.

The first step in the cancer registry process is case finding – identification of a reportable diagnosis of cancer – and one that has traditionally been performed manually by registry personnel. While it has always been a challenge to separate the pertinent records needing review from the non-useful records that can be ignored, the expanded use of electronic health records (EHR) systems increases the number of reports that will be easily accessible to the registrar. Manually identifying the pertinent reports becomes more time consuming due to this increased volume. An automated method for identifying relevant reports is needed to increase the efficiency and accuracy of the case finding process.

Purpose of Project

The Text Mining Tool Evaluation project will describe the process of text mining, identify non-proprietary software that can search blocks of text to identify reports relevant to the cancer registry, and provide information to state cancer registries regarding different tools available and a comparison of the functionality provided by

³ Ibid.

⁴ American College of Surgeons. Commission on Cancer: Cancer Programs Standards, Revised Edition. 2007.

each tool. Evaluating the ability of a tool to map relevant data to cancer registry structured data may also be explored.

Contents of this Report:

This report provides information on Phase I of the Text Mining Tool Evaluation

Project: describing the text mining process and identifying tools that may be useful within the cancer registry community.

Methodology

To identify and recommend an open source text mining tool, the evaluation team conducted a literature search as well as an evaluation of identified software tools.

The literature search included:

- 1) **Text mining textbooks** to gain an understanding of the concepts and terminology used in text mining.
- 2) **Online research articles** using Google Scholar⁵ and PubMed⁶ to identify concepts and methodology and to evaluate text mining results using open source software;
- 3) **Proprietary text mining white paper discussions** to identify common features and proprietary features that may be helpful in the cancer registry community.

⁵ <http://scholar.google.com/>

⁶ <http://www.ncbi.nlm.nih.gov/sites/entrez/>

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

Identified tools were reviewed to determine whether they would meet the needs of the cancer registration community.

Understanding Text Mining

Information Extraction and Information Retrieval⁷

There are two processes involved in text mining. *Information retrieval(IR)* involves finding documents that provide the information needed using indices; these are usually called search engines. IR returns documents, classifies documents as relevant or not relevant and doesn't care about syntax. *Information extraction (IE)* is used to extract information from text without requiring the end user to read the text. IE returns facts using natural processing language NPL and is based on syntactic and semantic analysis.

Information retrieval has been systematized and formalized for many years.

Information extraction, or text mining, is rapidly becoming a well established science. The IE field is currently:

- Standardizing IE terminology and definitions;
- Identifying the steps needed to fully retrieve knowledge from unstructured text;

⁷ *The definitions used in this Progress Report are taken from Ananiadou S, McNaught J., Editors. Text Mining for Biology and Biomedicine. Archtech House. 2006. The citation from their work should be used, rather than this internal document.*

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

Evaluation of Text Mining Tools

- Developing open source modules for performing the steps;
- Developing standards for acceptable text mining outcomes.

Definitions and Uses⁸

Text mining is not performed in one step, but rather in a series of modules that identify, categorize and document text so that it can be evaluated using standardized techniques. There are three main considerations:

- 1) **Lexical level:** Evaluating the words;
- 2) **Syntactic level:** Evaluating the organization of groups of words in sentences into phrases or clauses (units);
- 3) **Semantic level:** Evaluating the meaning that can be given to these units in terms of content.

Lexical Level

Lexical level processing identifies how each **word** should be identified and used.

Several methods are available:

⁸ *The definitions used in this Progress Report are taken from Ananiadou S, McNaught J., Editors. Text Mining for Biology and Biomedicine. Archtech House. 2006. The citation from their work should be used, rather than this internal document.*

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

Tokenization – Deciding what is a word, abbreviation, number, symbol or punctuation. The simplest method of tokenizing is to use white space and punctuation; however, this causes some problems⁹.

- 1) Does a period after a letter indicate the end of a sentence or that the letter is an abbreviation?
- 2) Is an apostrophe denoting contraction of two words (don't for do not) or is it expressing possession? Use of possessives can be very important for discovering relationships.
- 3) Are hyphens and numbers represented consistently amongst documents? (CIN II – III versus CIN 2, 3?)

In biology, and specifically cancer, abbreviations and punctuation can cause a sentence to have more than one meaning. For example: “ALL” can mean either “ALL specimens were negative” or “acute lymphocytic leukemia”.

Morphological Analysis identifies and assigns different forms of a word to the same base word. Nouns and verbs frequently have different forms, i.e. past, present and future verb tenses. These variations are either inflection (activate, activates, activated) or derivation (activating, activation). To perform morphological analysis just on words alone (no special content meaning), the most frequent method is the Porter algorithm¹⁰. It uses a set of suffixes such as es, ed, ing, ion, ly, and iteratively matches a word string from right to left based on the

⁹ Ibid.

¹⁰ Ibid.

longest match. It strips off the suffixes, and leaves what is assumed to be the canonical (most basic) word root.

Linguistic Lexicons are combination of the lexical (word) element, either as the full word or its canonical base form, together with additional information which is needed for further morphological, syntactic and semantic processing.

Parts of Speech (POS) tagging is an example of the information that is included with the word. Activate (POS Verb), Activation (POS Noun), and active (POS Adjective). This is discussed further under the Syntactic level of text mining modules. A second type of information can also be included, such as singular or plural, or past, present or future tense. A third type of information to provide semantic information is available for further classification.

Syntactic Level

Syntactical level processing splits groups of words in sentences into phrases or clauses (units).

Grammars are linguistic descriptions, usually in the form of rules or constraints which characterize well formed conditions (POS tags and features, noun phrases, prepositional phrases). Some text mining applications approximate the grammatical regularities using ad hoc pattern matching rules. These quickly achieve limited benefits, but usually fail to scale up for large and diverse document

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

collections because their rule specifications tend to become increasingly complex and harder to control and maintain.

Treebanks consist of a corpora (collection) of plain text, for which human annotators have supplied phrase annotations at the sentence level. Treebanks require a clear commitment to grammar theory and the capability of the annotators to understand the contents of the underlying documents. The advantage of this work is that the grammar rules don't have to be specifically listed, but can be learned automatically from the positive examples. Large volumes of text should be annotated by several (3+) human annotators to ensure consistency and quality. In the biology domain, GENIA TreeBank has POS annotations sufficient for learning purposes. Syntactic annotations are still under development.

Parts-of-Speech Tagging removes ambiguities in words according to the context of the phrase. A good example is to tag the word "report" to clarify its meaning. Is it a noun – a pathology report; or is it a verb – report the cancer diagnosis? There are rules-based taggers, where a set of rules are established and run against the corpora. The annotator reviews the rules result against the gold standard of the corpora, determines where errors were made and corrects the rules as needed.

Then the rules are assigned an application order. There are also statistical

taggers.¹¹ All subsequent modules for syntactical processing rely on the tagger's output. High performance of the tagger is critical for success in later stages.¹²

Chunking identifies discrete units of a phrase, such as noun, preposition, verb or adjective phrases. There are two types of chunking – *base NP chunking* (non-recursive noun phrases) and *text chunking*. Both rely on annotated corpora of phrase chunks for training.

Parsing identifies word sentences that contain a subject and a predicate, called clauses.

Semantic Level

Semantic level processing identifies the context in which phrases are used, for example, whether a sentence refers to an action happening to something or to the action itself [the diagnosis *of leukemia* or to *the process of diagnosis* itself].

Lexical and syntactic level processing are language-specific, dealing with parts of speech as they relate to general language. Semantic level processing is language-neutral, but has domain-specific concepts; in other words, semantic processing describes the relationship between words and phrases.

¹¹ Ibid.

¹² Ibid.

The text mining community divides semantic processing into two groups: terminologies and ontologies. A terminology provides a list of entities and link synonyms together. An ontology connects entities by documenting relationships between them. Rather than focusing on the names, an ontology defines domain classes and their inter-relationships. Because terminologies provide some hierarchy if only for organizational purposes, and because ontologies also collect names for their entities, the dividing line between the two is somewhat arbitrary.¹³ The main distinction as proposed by S. Ananiadou and J McNaught is that when relationship of the concepts is left implicit (the human user provides the relationships) it is called a terminology; if the relationships are formalized so that inferences can be automatically drawn they are ontologies.¹⁴

Lexicons, terminologies and ontologies are resources used in text mining to support entity recognition (finding terms that are of interest to the domain) and entity relationships. Entity recognition usually creates its list of entity names from various disease resources (ICD9, ICD-O, CPT, etc). Relation extraction usually comes from structured terminologies or ontologies. A lexicon will include words and multiword expressions that are frequently observed in the text corpora of a domain and record information about them, including parts of speech (noun,

¹³ Ibid

¹⁴ Ananiadou S, McNaught J., Editors. Text Mining for Biology and Biomedicine. Archtech House. 2006, page 26.

adjective), inflectional variants (singular, plural) and spelling variants (British and American).

Challenges in Identifying Terms

Term Variation

Term variations where upper or lower case doesn't matter, or where use of hyphens or parentheses are optional are called non-contrastive. These variations still represent the same preferred term. A good example is recording the variations in recording CIN. It can be CIN 3 or CIN III. However, term variations of "edge effect" are contrastive; they change the meaning of the term. An example of "edge effect" is the recording of a number at the position of the last word: grade 3 versus grade 4. The most frequent term variations are punctuation (bmp-4 and bmp4) and use of different numerals (CIN 2-3, CIN 2,3).¹⁵ Table 1 describes types of variation that affect text mining.

Orthographic	Variation in using hyphens, slashes, lower and upper case, spelling, etc.
Morphologic	Variation in inflection; plural or possessive forms
Lexical	True synonyms, e.g., heart attack and myocardial infarction
Structural	Permutations (integrin alpha 4 and alpha4 integrin)

¹⁵ Ibid.

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

	Using prepositions to indicate possessive nouns, e.g., cancer of brain and brain cancer Permutations of prepositions (cells from blood and cells in blood)
--	--

Acronyms and Abbreviations and Ambiguity

It is essential to identify acronyms and abbreviations and relate them to their expanded form. The process is difficult, especially as some acronyms are synonymous—the same term has more than one acronym— and others are ambiguous—the same acronym may correspond to different terms. Ambiguity can be a problem in text mining as a word or term can have many meanings. For example, a *pound* can relate to a weight or to a currency.

The Text Mining Process

A comprehensive discussion of the process can be found on pages 31–34 of Text Mining for Biology and Biomedicine.¹⁶ Table 2 provides a summary of the process.

Table 2: Summary of the Text Mining Process		
Step	Main Step	Sub Steps
1	Obtain a large collection of raw documents	<ul style="list-style-type: none"> Cleanse to get rid of text formatting code (RTF, HTML, PDF, etc)
2	Perform Lexical work	<ul style="list-style-type: none"> Split into tokens (words) by using a tokenizer

¹⁶ Ibid. pages 31 – 34.

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

Evaluation of Text Mining Tools

		<ul style="list-style-type: none"> • Parts of Speech (POS) tags to text tokens. Can also include <ul style="list-style-type: none"> ○ a named entity recognizer for domain-specific types of words (numerical strings (grade 3), ○ measurement units (2mm lesion) disease names (adenocarcinoma) and ○ Acronym detection (ALL=acute lymphocytic leukemia)
3	Perform Syntactic tasks	<ul style="list-style-type: none"> • Create phrase chunks by grouping POS tags into plausible composite units (Splitting a complex sentence into sequences of phrases.)
		<ul style="list-style-type: none"> • A parser may either relate these chunks according to grammatical criteria (the difference between the subject and object of the sentence) or assign additional internal syntactic structures to the chunks.
4	Perform Semantic Tasks	<ul style="list-style-type: none"> • Create relationships between words and phrases. <ul style="list-style-type: none"> ○ Use a terminology or ontology to link concepts and/or biologic terms and provide a relationship.
		<ul style="list-style-type: none"> • Map the tagged words to the concept level to determine if it is the predicate or the object of the predicate (usually a lexicon look-up). • check syntactic results to see which chunk/parse unit denotes a particular argument of the predicate (using a semantic role labeler—mapping rules similar to human-made grammars and coding rules)
5	Assess by a relevancy filter	<ul style="list-style-type: none"> • This step is frequently and unfortunately omitted and all results are merely passed on to the end-user, which overloads them with non-relevant reports.

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

Results of Literature Search

Most open source text mining efforts arise from work performed by the University of Sheffield, UK, and the Centre for Text Mining. The Centre developed an open source Natural Language Processor (NLP) framework called GATE – General Architecture for Text Engineering.^{17, 18} GATE includes a set of modules collectively called CREOLE (Collection of RE-usable Objects for Language Engineering) to provide a useable architecture and a suite of modules to perform text mining.¹⁹ GATE also provides various services to the modules, such as component discovery, bootstrapping, loading and reloading, management and visualization of the data structure and data storage and process execution.²⁰ GATE and CREOLE has become the de facto foundation within many software tools.

Cancer registries can use one or more of the CREOLE modules to improve the sensitivity and specificity of their text mining tools. Of immediate use is the negation module called Negex-2^{21, 22} which determines whether a diagnosis, for example, is present or absent. Negex is discussed later in this report.

¹⁷ Cunningham, H, Maynard, D., Bontcheva, K., Tablan, V. GATE: an Architecture for Development of Robust HLT Applications

¹⁸ <http://gate.ac.uk>

¹⁹ GATE: an Architecture for Development of Robust HLT Applications: <http://gate.ac.uk/sale/acl02/acl-main.pdf>

²⁰ Zeng, QT., et al. Extracting Principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Medical Informatics and Decision Making 2006, 6:30; July 26, 2006.

²¹ Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct;34(5):301-10

²² NEgex-2 can be found at: <http://web.cbmi.pitt.edu/~chapman/NegEx.html>

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

Open Source Tools²³

CaFE: Registry Case Finding Engine ²⁴
Responsible organization: University of Michigan USA
Purpose: Automated approach for cancer patient identification from unstructured, free-text pathology reports.
Method: Java Server Pages based application
Cancer Registry Use: Used in a hospital cancer registry, caFE appears to be a standard search term look up list that will highlight positive words/phrases and negative terms/phrases for review and to ignore other word/phrases of no interest. Depending on its handling of the negative terms (can it successful identify reports that truly have only negative cancer findings so that the report does not require manual review) it may be sufficient for central registry needs.
Reference: <i>Journal of Clinical Oncology</i> , 2006 ASCO Annual Meeting Proceedings Part I. Vol 24, No. 18S (June 20 Supplement), 2006: 6080

²³ Software tools discussed in this report do not constitute an endorsement by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for use of the mentioned software tools. This report provides a review of the mentioned software tools as a service to our partners.

²⁴ Danauer, D., Chinnaiyan, . Registry Case Finding Engine (caFE): An Informatics Tool to Identify Cancer Patients in Electronic Pathology Reports. *Frontiers in Oncology and Pathology*. Vancouver, BC, Canada, August 2006.

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

Extended MedLEE: Medical Language Extraction and Encoding System ²⁵
Responsible organization: Columbia-Presbyterian Medical Center
Purpose: Automated encoding of the information content of text documents, including discharge summary, radiology, pathology and mammogram.
Method: Perl preprocessor to put the reports into an eXtensible Markup Language (XML) version that MedLEE can analyze; contains specific text-mining modules to perform text analysis.
Cancer Registry Use: Can be used for identifying reportable diagnoses in many types of medical reports. Can also handle tabular type text reporting (i.e. “Her-2 score is 1+”), including the ability to add rules to interpret and recode discrete data values to the registry standard data values. (“Her-2 score of 1+ is recorded in the database as Her-2 <i>negative</i> .)
Reference: <i>Automated Encoding of Clinical Documents Based on Natural Language Processing</i> http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=516246 <i>Facilitating Cancer Research using Natural Language Processing of Pathology Reports</i> http://cmbi.bjmu.edu.cn/news/report/2004/medinfo2004/pdf/papers/4468Xu.pdf

²⁵ Xu, H., Anderson, K., Grann, V., Friedman, C. Facilitating Cancer Research using Natural Language Processing of Pathology reports. MEDINFO 2004.

HITEx: Health Information Text Extraction
Responsible organization: Harvard University, USA
Purpose: An opens source full text mining software package.
Method: Using GATE as a platform, a suite of open-source NLP modules were adapted or created. HITEx assembles these modules into pipelines for different tasks. ²⁶
Cancer Registry Use: Tested using discharge summaries, which are less standard than pathology reports and contain extraneous medical information and diagnoses that may or may not be of interest. HITEx tested well against identifying principal diagnosis and co-morbidity. It has an ability to differentiate between personal and family history, and to extract temporal modifiers within a discharge summary.
Reference: <i>Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system.</i> http://www.biomedcentral.com/1472-6947/6/30

²⁶ Zeng, QT., etal. Extracting Principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Medical Informatics and Decision Making 2006, 6:30; July 26, 2006.

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

caTIES: cancer Text Information Extraction System
Responsible organization: National Cancer Institute, USA – caBIG Program
Purpose: A general purpose text information extraction tool to automate the process of converting free text pathology reports into structured data, storing and facilitating advanced query and analysis of the pathology information
Method: Uses GATE as a platform. Spin pipeline uses: DeID, Tokenizer, G Spell, Chunker, Concept Tagger (UMLS) NegEx, Concept Mapper, Thematic Role Mapper, XMLizer. ²⁷
Cancer Registry Use: Contains a comprehensive, fully integrated package for text mining pathology reports. caTIES may include CREOLE modules that are well beyond registry needs. This is not important if the tool is easily implemented, but may be a concern if manual results processing is more than expected or final result outcomes are less than expected.
Reference: http://caties.cabig.upmc.edu/

Proprietary Software

A website review of 3 proprietary tools—Semantic-Knowledge’s Semantic Engine, Attensity’s Text Analytics Suite, and SAS’s Text Miner—was performed to

²⁷ See website: <http://gate.ac.uk/sale/acl02/acl-main.pdf> for description of modules.

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

determine whether features useful to the cancer registry community were available in proprietary software that have yet to be included in open source software.

Discussion

Tool Evaluation

caFE

An evaluation copy of caFE was not available due to university policy. Discussions with Dr. Hanauer indicate that caFE is at a lower level of complexity as HL7 MapperPlus, a tool developed within NPCR to parse HL7 cancer data into database format, filter out non-relevant records, and present a user interface for registrars to review records. It uses a different search term list; efforts are underway to get approval to compare the caFE search term list with the NAACCR search term list to identify differences that could help both lists.

caFE'S level of sensitivity is higher than those obtained by using the NAACCR search term list, which may be due to selection bias. caFE processes inpatient reports from the University of Michigan hospital, thus enabling more traditional text diagnoses of cancer than the reference laboratory serving as the NPCR-AERRO ePath Pilot Project's laboratory.

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

A website has been established to evaluate caFE.

<http://bioinformatics.cancer.med.umich.edu/tools/casefinding/>

MedLEE

MedLEE is a full text mining solution which may have utility in the cancer registry community. While MedLEE is a freely available tool, Columbia University has a formal process for obtaining a license for using its software. The License Agreement requires review by institution legal or general counsel's office and signoff by an executive officer who is authorized to enter into legal agreements on behalf of the institution.

It also appears that MedLEE maintains licensing control and each registry would need to apply for a license. With the availability of two other freely available, non-licensed products, further exploration of this tool was deferred until the results of evaluating the other tools is completed.

HITex

HITex uses GATE and java and is a full text mining solution. It may be more complex than what is required for electronic pathology report mining. Modules

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

could be extracted as stand-alone components. The regular expression extraction tool uses the Unified Medical Language System UMLS semantic types. It would need to be modified to use the North American Association of Central Cancer Registries (NAACCR) Search Term List. The negation module would probably require the most modification.

To minimize the likelihood of finding a positive match between the Search Term list and information in the Clinical History (or other) section of the report, the research suggested using the sectionizer module. The methodology is to prepare a configuration file for each report type that includes the section headings used within the report.

HL7 MapperPlus could use this methodology by building configuration files for each laboratory and their associated reports. For example, LabCorp submits four types of reports to central registries: Surgical Pathology, Fine Needle Aspirates, Non-GYN Cytology and GYN Cytology. A configuration file for each of these reports, with LabCorp section headings, could be referenced to select which section of the report should be scanned with the Search Term List. Performing this task would minimize false positives due to past history of diagnoses documented in the clinical history, or possible diagnoses of cancers listed in the frozen section diagnosis section.

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

Evaluation of Text Mining Tools

The development contact recommended having the NPCR-AERRO programmer review the documentation and source code and send questions via email. The developers are available to help explain the tool and offer suggestions on how it can be adapted for NPCR-AERRO use.

ca_TIES

ca-TIES is the most comprehensive freely available text mining solution. It was developed specifically for mining text within pathology reports, so it has a one to one objective with the NPCR-AERRO ePath Pilot Project needs. ca_TIES is part of the ca-BIG project by the National Cancer Institute (NCI) and is interoperable with many other NCI resources (National Library of Medicine's UMLS, etc). Its use will provide another tie between the NCI and CDC that will support the goal of cooperation and collaboration.

ca_TIES is an open source and freely available complete text mining tool, providing a comprehensive suite of modules for text mining. It has modules that stand independently so that NPCR can select those that are needed; its modules can be modified by the NPCR to meet any project specific needs. ca_TIES is fully supported, maintained and enhanced using a team approach to ensure that it maintains its high quality IT function and continues to meet the needs of the end user. ca_TIES may be more tool than cancer registries need for text mining of

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

pathology reports. A production version is currently available for download and evaluation.

Conclusions

Text mining can be an effective tool for cancer registries to assist in identifying essential concepts and data that are required in a cancer case summary, and to provide concepts of secondary use that aren't traditionally included due to the amount of resources required to collect the data.

Text mining of pathology reports can go beyond identifying reportability and the histologic type of cancer. Collaborative stage data items, such as tumor size and CS site-specific factors can be collected easily. From the patient history and physical, text mining can identify race, family history, and smoking history. The discharge summary can be mined to obtain the treatment plan and co-morbid conditions.

Most proprietary text mining companies highlight the scalability, management of annotations/lexicons, administration and security features within their software. Each claims to use unique concepts and/or methods of analyzing text, leading to a competitive edge over the other tools available. The feature most frequently mentioned that does not appear to be available in open source software is an end

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

Evaluation of Text Mining Tools

user interface that allows question-answer sessions after the unstructured text has been machine analyzed. Unless the open source tools identified fail in sensitivity and specificity, proprietary text mining tools need not be considered.

Evaluation of the basic components of text mining revealed that use of GATE-CREOLE's NegEx, a freestanding open source module for determining negation, could greatly improve the specificity of current text mining activities in cancer registration. NegEx was evaluated and implemented into the NPCR's HL7MapperPlus program. A trained set of 50,000 electronic pathology reports from the Florida Cancer Data System that were reviewed and coded by certified tumor registrars confirmed the utility of providing a negation function to minimize false positives and are described in a separate report.²⁸ Negex has been incorporated in the NPCR-AERRO HL7 MapperPlus tool to identify negated cancer terms within electronic pathology reports. HL7 MapperPlus can be downloaded from <http://www.cdc.gov/cancer/npcr/tools/>.

Extended MedLEE, HITEx, and caFE were developed to solve a need within a specific medical institution. Preprocessing work to split the reports based on institution-specific recording standards may contribute to a more favorable true-positive and true-negative result than what may be obtained outside the institutions. Neither Extended MedLEE nor caFE is fully portable and will need more programming support to build a tool that is portable to all NPCR registries.

²⁸To be submitted to a cancer registry peer-reviewed journal.

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

HITex and caTIES are fully portable, but may be more tool than is needed for text mining of pathology reports. HITex has an advantage over ca-TIES in that it includes text mining of the more complex discharge summaries as part of its capabilities, which could be useful when other EHR reports are submitted to cancer registries. Alternately, ca-TIES has strong interconnectivity to other tools used in the cancer registry community and also has the infrastructure to provide support to NPCR-AERRO efforts.

A registry or vendor wishing to build its own text mining tool should fully evaluate GATE before initiating their own design. The modules will perform each task; however, software will need to be written to tie them together into a logical sequence of processing.

Action Plan

Additional work will be undertaken to expand text mining efforts in cancer surveillance. NPCR-AERRO plans to evaluate Ca-TIES to determine the usefulness of each module to meet the cancer registry text mining needs. NPCR-AERRO also plans to evaluate HITex to determine its usefulness for reports such

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.

Evaluation of Text Mining Tools

as discharge summaries, etc, where there is less formal and consistent formatting of the report than in either pathology or radiology reports.²⁹ Evaluation will include:

- Ease of installation;
- Plug and play ability;
- Completeness of documentation;
- The need for and the ability to modify the tool to meet cancer registry needs;
- Evaluating the performance of the software tool against a trained set of electronic pathology reports, and against discharge summaries and clinician notes.

The results of this in depth evaluation will be reported in a second report and will be posted on the NPCR-AERRO website at

<http://www.cdc.gov/cancer/npcr/informatics/aerro/>.

²⁹ NPCR-AERRO's intent to evaluate ca_TIES and HITex does not constitute an endorsement by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for use of the mentioned software tools. This evaluation provides a review of the mentioned software tools as a service to our partners.

Links to non-Federal organizations are provided solely as a service to our users. Links do not constitute an endorsement of any organization by CDC or the Federal Government, and none should be inferred. The CDC is not responsible for the content of the individual organization Web pages found at this link. All links were active at the time of publication.