Concept Discovery for Pathology Reports using an N-gram Model

Vincent Yip[1], Mutlu Mete[2], Umit Topaloglu[1], Sinan Kockara[3]
[1]University of Arkansas for Medical Sciences, [2]Texas A&M University-Commerce,
[3]University of Central Arkansas

Abstract

A large amount of valuable information is available in plain text clinical reports. New techniques and technologies are applied to extract information from these reports. One of the leading systems in the cancer community is the Cancer Text Information Extraction System (caTIES), which was developed with caBIG-compliant data structures. caTIES embedded two key components for extracting data: MMTx and GATE. In this paper, an n-gram based framework is proven to be capable of discovering concepts from text reports. MetaMap is used to map medical terms to the National Cancer Institute (NCI) Metathesaurus and the Unified Medical Language System (UMLS) Metathesaurus for verifying legitimate medical data. The final concepts from our framework and caTIES are weighted based on our scoring model. The scores show that, on average, our framework scores higher than caTIES on 848 (36.9%) of reports. Furthermore, 1388 (60.5%) of reports have similar performances on both systems.

1. Introduction

Nowadays, an ever changing world of technology produces a vast amount of information in different fields. Likewise, pathological data are part of this ocean with much valuable information. The challenge includes complex systems that would provide proper data to physicians on demand basis for quality patient care. To enable a pathologist being able to quickly identify data from massive information, accurate computer-assisted decision support systems with text mining abilities are crucial [1].

In addition, from initial diagnosis to definitive treatment, pathology evaluations play an important role in the cancer patient care. Since most patient management depends on the right biospecimen diagnosis, the pathology stage is widely considered the most accurate predictor of survival. It also determines the appropriateness of adjuvant treatment. Various additional pathology factors have been shown by multivariate analysis to have prognostic significance that is independent of stage. These may help to further sub-stratify tumors, individualize treatment, and more accurately predict outcome. On a larger scale, pathology data are essential for epidemiology and clinical research. Therefore, it is known as the common language of cancer worldwide [2].

Since the data embedded in pathology reports are so valuable, concepts have to be extracted accurately. Furthermore, information needs to be discovered with text mining techniques before the data becomes accessible by physicians. In an attempt to overcome this challenge, an n-gram based text mining approach is adopted to extract valuable concepts from pathology reports. Different technologies and methods are reported in the literature in order to extract data from varies medical reports [3-10].

In this study, an n-gram algorithm is used to find the common theme, concepts, in pathology reports. A word or a group of consecutive words that occurs frequently enough in the entire report collection is considered as a concept. Each concept candidate is expected to fulfill a predefined frequency threshold in order to become a concept. The frequency threshold is explained in section 3.1. N-gram algorithm is chosen because it is domain independent [11], unlike Weeber et al. [12], who mapped sentences to predefined UMLS [13] concepts. In our study, the UMLS and NCI Metathesaurus are only used for filtering our results for scoring purposes.

2. Resources

Two sources of vocabulary knowledge were used by this study: the UMLS and the NCI [14] Metathesaurus. The UMLS Metathesaurus is the foundational knowledge, which is the base of the comprehensive thesaurus and ontology of biomedical concepts. It consists of a collection of terms from different controlled vocabularies and their relationships. The NCI Metathesaurus is based on the UMLS Metathesaurus; however, it is supplemented with additional cancer-centric vocabulary.

In this study, the MetaMap online tool, which accesses the UMLS and the NCI Metathesaurus, is used exclusively. MetaMap returns relevant concepts based on the UMLS and the NCI Metathesaurus when a possible concept is provided. A concept list that is generated from our system is passed to MetaMap for scoring. Only the concept score of 1000 is granted for exact concept matches. These exact matches become our final concept candidates for reconsideration in our framework. Other non-exact matched concepts are marked with lower scores such as 900, 800, etc. MMTx [15] is a java implementation of MetaMap and is used by caTIES. In this study, MetaMap is used with the default options. Note that since concepts are validated by MetaMap and only exact matches are used, concept overlapping is not considered.

caTIES [16] is a silver level caBIG-compliant open source text extraction system. Legacy, Bronze, Silver, and Gold level compatibility represents tool's ability to interoperate with other systems and assigned after the caBIG review process. The Cancer Biomedical Information Grid (caBIG) [17] is an initiative of the NCI, which is a part of the National Institutes of Health. It is a truly collaborative information network for cancer researchers, to share knowledge and data. caBIG enables and encourages the discovery of new ideas for the detection, treatment, diagnosis, and prevention of cancer in order for the cancer community to improve patient outcomes. One of the chief strengths of caBIG is its ability to join research tools, data, scientists and the cancer community. This combined strength and expertise in an open environment is the mission of caBIG.

In this study, caTIES is used as a control system. It extracts coded information from free text pathology reports using varies natural language processing (NLP) techniques. GATE (General Architecture for Text Engineering) [18] is the main part of the NLP core of caTIES and is used extensively. GATE is a java toolkit for NLP. By using some publicly available NLP tools, algorithms, and the NCI Metathesaurus, caTIES is capable of identifying and indexing concepts from pathology reports.

2.1 Dataset

The most recent two weeks' surgical pathology reports (total 2,295) were obtained from the University of Arkansas for Medical Sciences database as our dataset. They were selected from a fixed time range (between 6/22/09 and 7/6/09). These reports have an average of 151 words. Among them, 19% are surgical report, 18% are dermatopathology report, and 11% are cytogenetics report, etc. These reports are the most frequent in the dataset.

3. Methodology

In this section, we discuss how we process the text report and extract concepts from our system. The next section presents how legitimate concepts are processed and verified. Finally, we introduce a concept scoring model to rank our system against caTIES.

3.1. Data Extraction with the n-gram Approach

Our model consists of three main components: a non-character filter, a stop word filter, and an n-gram generator. The non-character filter removes non-characters from all reports including double spaces, numbers, and punctuation, etc. Double spaces are replaced with a single space. This ensures that empty spaces will not be treated as part of a concept. At this stage, numbers are not considered as part of a concept. In addition, stop words (such as a, an, and the, etc.) are removed since they are not part of medical concepts. In this study, caTIES's stop word list is used. In our n-gram algorithm there are two main parameters: maximum number of grams (MNOG) and frequency. In our experiment, MNOG and frequency range from 3 to 5 and from 3 to 10 respectively.

The MNOG defines the maximum number of words that a concept should consist of. For instance, if MNOG is set to four, only concepts with at most four words are visible e.g., "Left breast cancer cell" is considered as a concept whereas "Left breast cancer cell shows red spots" is not considered as a concept. Instead, "Left breast cancer cell shows red spots" are two concepts: “Left breast cancer cell” and “red spots”. MNOG is also one of the crucial parameters in our model. In case a dataset includes a number of 4-gram concepts (i.e., concepts that are four words long) and MNOG is set to 2, then these concepts are divided into two separate parts. Therefore, using a smaller number for MNOG tends to both lose actual concepts, and unnecessarily increase the number of shorter concepts.

The frequency controls how frequent a concept candidate should appear among all reports. For instance, if “breast cancer” appears ten times within all reports, while the frequency is set to five, “breast cancer” are considered as one of our concept candidates. However, if a term occurs once in all reports, this term is treated as a non-significant medical related term. Therefore, the frequency control enforces differentiating medical terms from everyday words while keeping frequently used terms together.

In order to obtain concepts, our algorithm performs two major steps: generating candidate concepts and validating candidates based on the frequency.

Higher order n-grams, 5-gram, are generated first so that it will not split words apart from their neighbors (consecutive words). For each ‘n’, where ‘n’ is the number of words in the concept, the algorithm passes through the data collection once.

Once a list of concept candidates is generated, the frequency is used to check against the concept candidate list. Those concepts, which satisfy the frequency threshold, are considered as active concepts. In order to prevent concept reconsideration, as mentioned in section 2, these active concepts are removed from the data collection. In addition, candidate concepts generation and validation are processed for each gram.

3.2. caTIES Data Extraction

The same pathology reports are passed into caTIES for concept coding. Concepts of caTIES are stored in a centralized database in compressed binary format. In this study, they are decoded and stored in a concept list. Some ‘exact duplicate’ concepts were removed from the list.

3.3. Legitimate Concept Validation

MetaMap batch online tool is used to validate concepts from both our system and caTIES. Two separate lists were generated after data extraction with both our approach and caTIES. Then, these lists are passed into MetaMap. MetaMap provides scores for all concepts and is based on the UMLS and NCI Metathesaurus. If there is an exact match being found, the score is 1000. In this study, only the exact matching results are considered in order to simplify our comparison. After all the concepts are being evaluated, a list of concept scores for our system and caTIES are generated.

3.4. Legitimate Concept Processing

The list of exact match concepts both for our system and caTIES are being counted from the reports. Those concepts that were counted are completely removed from the dataset in order to avoid concept recounting. Higher gram concepts are considered first so that longer concepts are preserved. Therefore, the number of concepts that each system recognizes are recorded. As a result, comparisons of both our system and caTIES become possible by using our scoring model.

3.5. Concept Scoring Model

If a system discovers a concept, for example “Colon Cancer Treatment” while another system found “Colon Cancer” and “Treatment” separately from the same report, a method is needed to determine which system is more accurate. In most cases, “Colon Cancer Treatment” should be one concept instead of two. With this philosophy in mind, a concept scoring model is developed to rank the performance of our system. The total concept score for a report is denoted as (ξ).

$$\xi = \begin{cases} \sum_{p=1}^{L} \delta\, Kt, & if\ L > 0 \\ 0 & ,\ if\ L = 0 \end{cases} \quad K = \begin{cases} 1\ if\ \delta = 1 \\ 2\ if\ \delta > 1 \end{cases} \text{(eq. 1)}$$

In equation 1, L represents the total number of concepts in a report. If no concepts are found (L=0), ξ is zero. p represents the index of concepts, t is the concept occurrence in a report, and δ is the number of grams of a concept. K is a constant and its value depends on δ. Assuming that we found L number of concepts (C_1, C_2... C_L) in the Nth document in reports (R_1, R_2... R_N). If the number of grams of C_1 (p=1) is one ($\delta=1$), then K is set to 1, otherwise 2. This is because higher order concepts are more important than 1-gram concepts. The concept score (ξ) depends linearly on t because concepts with higher frequency in the report should be favored.

In section 3.3, an individual concept list is generated for both systems pertaining to each report. The scoring model is then applied to these lists to calculate how each system scores on each report. One point is added to a particular system if it is determined that it scores better on a report than the other system. If our system scores the same or better than caTIES, it demonstrates that our approach is capable of extracting valid medical terms from pathology reports.

4. Experiments and Results

2,295 pathology reports were selected from our database to demonstrate efficiency and robustness of the proposed system. According to our experiment as shown in Table 1, the specification of the MNOG and the frequency affects the results significantly (As mentioned in section 3). In order to obtain the optimum results, nine parameter pairs were selected as shown in Table 1.

According to our results in Table 1, our system scores higher than caTIES on an average of 36.9% of reports. This percentage of documents generated higher scores based on our scoring model. On the other hand, both systems have similar performance on an average of 60.5% of reports. A time-wise comparison will be one of our future works. Once concept scores are assigned to each system for each report, the concept score difference (**Φ**) is found. Thus, three result conditions are obtained: (a) tie (where **Φ** <= 10), (b) lose (caTIES performs better), and (c) win (n-gram performs better). Since our largest MNOG is five and the maximum K value in eq. 1 is two, the highest single score increment is ten. Therefore, the concept score difference less than or equal to ten points is considered to be a tie.

This promising result shown in Table 1 indicates that our model is capable of effectively extracting concepts from pathology reports. The next challenge is: what parameter specifications should be used to obtain the most accurate results. From results in Table 1, it is observed that with the same gram settings; when the frequency (t) increases, q (the total number of reports with **Φ** greater than ten with the proposed algorithm) decreases. This suggests that t is inversely proportional to q ($t \sim {}^{1}/_{q}$) (Table 1).

Table 1. Score comparisons with different parameter specifications (sorted by MNOG)

Parameter Spec.		System	# of Reports	
MNOG	Freq.		Φ * <= 10	Φ * > 10
-	-	caTIES	1185	2
3	3	n-gram		**1108**
-	-	caTIES	1262	7
3	5	n-gram		**1026**
-	-	caTIES	1679	28
3	10	n-gram		**588**
-	-	caTIES	1150	1
4	3	n-gram		**1144**
-	-	caTIES	1328	58
4	5	n-gram		**909**
-	-	caTIES	1680	55
4	10	n-gram		**560**
-	-	caTIES	1346	2
5	3	n-gram		**947**
-	-	caTIES	1276	182
5	5	n-gram		**837**
-	-	caTIES	1588	195
5	10	n-gram		**512**
Average		caTIES	1388	59
		n-gram		**848**

* **Φ** is the score difference for each report between two sample systems.

Also, the relationship between the MNOG and q is realized. The MNOG and q are also inversely proportional ($\mathrm{MNOG} \sim {}^{1}/_{q}$) to each other. However, there is an exception: when q reaches its optimum result. This happens when the MNOG is set to 4 and the frequency is set to 3: our system scores higher than caTIES on 49.9% of reports and both systems have similar performances on 50.1% of reports.

One reason our system has scored better than caTIES is because our scoring model is being used. The scoring model is designed to favor a system that discovers higher gram concepts. Thus, a concept that is longer in length scores higher with our scoring model.

5. Discussion

One disadvantage our framework has is its dataset dependent nature. Therefore, our results are highly correlated to the data collection. For instance, a term only appears once in all reports, which is less than our frequency threshold, will not be considered as a concept. Since some specific terms will only appear in certain types of pathology reports, these terms will be missed by our system. One way to address this issue is to classify pathology reports by their type. Thus, the data collection size for different type of reports will be controlled. This ensures that enough training data for various types of pathology reports is obtained.

6. Conclusion and Future Work

In this study, our system scores higher than caTIES on an average of 36.9% of reports. On the other hand, both systems have similar performance on an average of 60.5% of reports. Although promising results are generated, there is still room for improvement. Some future work includes incorporating MetaMap with our algorithm. MetaMap can be used as a mean for suggesting and breaking down invalid concepts. In addition, numbers and symbols will also be taken into the consideration as part of concept candidates. This in turn will provide more information to MetaMap while scoring concepts. Moreover, our training dataset will be tailored based on the report type. This will increase the frequency of the legitimate concepts. In addition, the time-wise comparison will be evaluated in the future.

References

[1] Sinard JH, Morrow JS. Informatics and anatomic pathology: meeting challenges and charting the future. Hum Pathol 2001;32:143–148.

[2] Compton CC. Surgical pathology for the oncology patient in the age of standardization: of margins, micrometastasis, and molecular markers. Semin Radiat Oncol. 2003;13:382–388.

[3] W. Long, Extracting Diagnoses from Discharge Summaries, Proc AMIA Symp; 2005. pp. 470–474.

[4] W.W. Chapman, J.N. Dowling and M.M. Wagner, Fever detection from free-text clinical records for biosurveillance, J Biomed Inform; 2004. pp. 120–127.

[5] M.L. Morsch, J.L. Vengo, R.E. Sheffer and D.T. Heinze, CM-Extractor: An Application for Automating Medical Quality Measures Abstraction in a Hospital Setting, AAAI; 2006. pp. 1814–1821.

[6] C. Friedman, L. Shagina, Y. Lussier and G. Hripcsak, Automated encoding of clinical documents based on natural language processing, J Am Med Inform Assoc 11 (5); 2004. p. 392–402.

[7] Q. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. Murphy and R. Lazarus, Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Dec Mak 6; 2006. p.30.

[8] M. Sordo and Q. Zeng, On sample size and classification accuracy: a performance comparison Lecture Notes in Computer Science, 3745; 2005.

[9] Wilbur W.J.et al. Analysis of biomedical text for chemical names: A comparision of three methods. Proceedings of the 1999 AMIA Annual Fall Symposium; 1999. p.176–180.

[10] Wilbur,W.J. and Lipman,D.J. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl Acad. Sci. USA 1983;80: 726–730.

[11] Miller E., Shen D., Liu J., Nicholas C.. Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System. Journal of Digital Information 2000 Jan;1(5):1-25.

[12] Weeber M, Vos R, Klein H, de Jong-van den Berg LTW. Using Concepts In Literature-Based Discovery: Simulating Swanson'S Raynaud Fish Oil And Migraine Magnesium Discoveries. Journal of American Society for Information Science and Technology 2001;52(7) :548-557.

[13] UMLS Metathesaurus Fact Sheet. [Online]. Available at: URL:http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

[14] NCI Metathesaurus. [Online]. Available at: URL:http://ncim.nci.nih.gov/ncimbrowser

[15] MetaMap Transfer (MMTx) Home. [Online]. Available at: URL:http://ii-public.nlm.nih.gov/MMTx/

[16] Chavan G. caTIES: Home. [Online]. Available at: URL:http://caties.cabig.upmc.edu/

[17] About caBIG®. [Online]. Available at: URL:https://cabig.nci.nih.gov/overview/

[18] GATE: a full-lifecycle open source solution for text processing. [Online]. Available at: URL: http://gate.ac.uk/overview.html