MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Context-Sensitive Medical Information Retrieval

Mordechai Averbuch [a], Tom H. Karson [b], Benjamin Ben-Ami [c], Oded Maimon[d], Lior Rokach[d]

[a]*Tel-Aviv Sourasy Médical Center and Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel*
[b]*Departments of Clinical Informatics and Cardiology, Mount Sinai School of Medicine, New York, USA*
[c]*Department of Computer Science, Tel-Aviv University, Tel-Aviv, Israel*
[d]*Department of Industrial Engineering, Tel-Aviv University, Tel-Aviv, Israel*

Abstract

Substantial medical data such as pathology reports, operative reports, discharge summaries, and radiology reports are stored in textual form. Databases containing free-text medical narratives often need to be searched to find relevant information for clinical and research purposes. Terms that appear in these documents tend to appear in different contexts. The context of negation, a negative finding, is of special importance, since many of the most frequently described findings are those denied by the patient or subsequently "ruled out." Hence, when searching free-text narratives for patients with a certain medical condition, if negation is not taken into account, many of the retrieved documents will be irrelevant.
The purpose of this work is to develop a methodology for automated learning of negative context patterns in medical narratives and test the effect of context identification on the performance of medical information retrieval. The algorithm presented significantly improves the performance of information retrieval done on medical narratives. The precision improves from about 60%, when using context-insensitive retrieval, to nearly 100%. The impact on recall is only minor. In addition, context-sensitive queries enable the user to search for terms in ways not otherwise available.

Keywords:

Information Storage and Retrieval, Medical Informatics, Information Management, Information Systems

Introduction

In a 1973 review the Chief of the Computer Research Branch at the US National Institutes of Health asserted that the data underlying the patient care process "are in the large majority nonnumeric in form and are formulated almost exclusively within the constructs of natural language [1]." Today, over 30 years later, much of the data stored in hospital information systems are still stored as free-text, including history and physical exams, pathology reports, operative reports, discharge summaries, and radiology reports. Databases containing free-text medical narratives often need to be searched to find relevant information for clinical and research purposes.

Medical narratives present some unique problems. When a physician writes an encounter note, a highly telegraphic form of language may be used. There are often very few (if any) grammatically correct sentences. Acronyms and abbreviations are frequently used. Very few of these abbreviations and acronyms can be found in a dictionary and they are highly idiosyncratic to the medical domain and local practice. Often misspellings, errors in phraseology, and transcription errors are found in dictated reports.

Various articles have been published evaluating methodologies for efficient information retrieval in the medical domain [2][3][4]. A search for patients with a specific symptom or set of findings might result in numerous records retrieved. The mere presence of a search term in the text, however, does not imply that records retrieved are indeed relevant to the query. Depending upon the various contexts that a term might have, only a small portion of the retrieved records may actually be relevant.

A number of investigators have tried to cope with the problem of a negative context [5][6][7]. Their detection of negative context is based on a regular expression built from a short list of negative terms supplied by a human expert. There is no work that tries to learn the profile of a context automatically and then uses this profile to examine various methods of context classification in the medical domain. Moreover, no work has been done to measure the effect of context on the result of medical information retrieval. The purpose of this work was to develop a methodology for learning negative context patterns in medical narratives and measure the effect of context identification on the performance of medical information retrieval.

Methods

Overview

Figure 1 presents a block diagram of the different components of the system. All medical documents are loaded into a database. Human experts review each document. Using a context tagging application, the experts specify the context (c) of each appearance of a medical term (t). The set of available contexts (C), where $C=\{C_1,...,C_n\}$, is predefined based on the specific application. For instance, in negation detection [5] the context set is *C={"Negative", "Positive"}*.

The resulting context-tagged document dataset (D) is divided into 2 sets: (1) the *training set* which contains two-thirds (2/3) of the documents along with the context of a few of the medical terms and (2) the *test set* which contains the remaining documents along with the context of few different medical terms. The

training set and the test set, therefore, contain different documents and the context of different medical terms.

The training set serves as the input to the *learning algorithm*. The output of the learning algorithm is the *context profile (L)*. Each context has its own profile that consists of a list of indicative terms. For instance the profile of a negative context may be $L^{negative}$={*"negative for", "denies"*}.

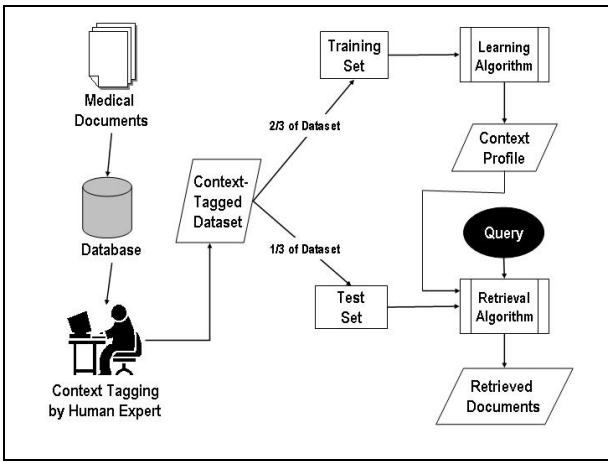

Figure 1 - Overview of Methodology

The context profile becomes an input into the *retrieval algorithm*. Queries for terms found in the test set are then created and run utilizing the retrieval algorithm, resulting in a set of *retrieved documents*. The recall, precision and F-measure [8] were measured for each of the queries.

Learning Algorithm

The core of the system is the learning algorithm. Its output, the context profile (L), is created by scanning the documents in the training set. All words or phrases (w) that appear in the same sentence as a tagged term are put on a list and statistics are generated regarding their appearance in other contexts. This list is then filtered using a threshold parameter, to eliminate rare words or phrases. Based on the UMLS Dictionary [9], the list is further reduced by removing all words or phrases that have medical context. (This removes medical terms that tend to correlate with tagged terms.) The next step is calculating the information gain (IG) for each term in each context. Equation 1 shows how IG is calculated, where $H(c)$ is the entropy of the context c and $H(c|term)$ is the conditional entropy for the context of the given term.

$$IG(c, term) = H(c) - H(c|term)$$

The last step of the algorithm is to remove terms from each context profile whose IG is below a given threshold. Pseudo-code of the algorithm is shown in Figure 2 .

Sentence Boundary Identification

The learning algorithm presented above requires that free-text be broken up into sentences. Normally, sentence boundaries can be detected by scanning the text for a period, exclamation point or question mark. This approach, however, does not work for medical narratives.

Input:
$C = \{c_1, ..., c_n\}$ – A set of Contexts (c_1 through c_n)
$D_{train} = \{d_1, ..., d_m\}$ – A set of Documents (d_1 through d_n); e.g., the Training Set
$T = \{w_1, w_2, ..., w_q\}$ – Dictionary terms (for instance the UMLS Dictionary)
$M = \{(d_i, t_{j_i}, c_l)\}$ – A set of manually tagged terms,
where c_l is the context of term t_{j_i} in document d_i
min_a – The value of the minimum required number of appearances
min_i – The minimum value of the threshold parameter for information gain (IG)
$L_t^c = \emptyset, \forall c, t$
For each $c \in C$
For each document $d \in D_{train}$
For each term, t s.t. $(d, t, c) \in M$
For each word (and phrase) w in the sentence of t
If $w \notin L_t^c$
Then add w to L_t^c
Else
Increase the number of appearances of w in L_t^c
Define $L^c = \bigcup_t L_t^c$ – A Context Profile
Remove from L^c all words with appearances of less than min_a
Remove all words in L^c that appear in T
For each $w \in L^c$ calculate the information gain $IG(c, w)$
Remove from L^c all words where $IG(c, w)$ is less than min_i
Return $L^{c_1}, ..., L^{c_n}$

Figure 2 - Pseudo-Code of the Proposed Learning Algorithm

As can be seen below, periods are frequently used within a sentence:

- Patient was discharged on Lopressor 25 milligrams p.o. b.i.d.
- After multiple attempts only 750 cc. of fluid were removed.
- He was evaluated by Dr. ___ of Neurology.
- Rechecked potassium was 4.4.

In this study sentence boundary determination still begins by scanning text for periods. Then, each period is evaluated to determine if it is part of a *regular expression*. Table 1 shows regular expressions written using Perl notation. If a period is part of a regular expression, it is marked as "not a separator." All other periods are considered sentence separators.

Table 1: Regular Expressions Marked As "Not a Separator"

(b|t|q)\.i\.d\.?	p\.o\.?	\.([0-9]+)	cc\.
p\.r\.n	q\.d\.?	\. of	\., and
q\.h\.s	mg\.	(Dr\.)(\s?)(\w+)	\sq\.

Retrieval Algorithm

The retrieval part of the experiment is meant to simulate queries made by physicians. All the documents in the test set are scanned for the query terms tested. In each document where query terms are found, a context classification, either positive or negative, is made for each appearance of the term. The context is classified by searching all the terms of the sentence where the query term is found and comparing it to the negative context profile. If a term is found in the negative context profile, that appearance of the query term was marked as negative. After classifying all appearances of the query terms in a document, the document is re-

trieved only if at least one appearance of the query term is in a 'positive' context.

Experimental Study

The potential of the proposed method for use in real word applications was studied. A database was created containing 4129 fully de-identified discharge summaries obtained from Mount Sinai Hospital in New York City. The database was divided into two groups using a 2:1 ratio. The *training set* consisted of 2752 documents (two-thirds of the total) and the *test set* contained 1377 documents.

Search terms chosen for retrieval in both datasets were: *nausea*, *abdominal pain, weight* loss and *tobacco*. In addition, the test set was searched for the terms: *headache, hypertension* and *chills*. This list of terms was chosen to represent different aspects of medical queries: simple terms (e.g., nausea), terms that are more than one word, popular terms, and terms measured with numerical values (e.g., 10 lbs weight loss).

The context tagging application highlighted each appearance of the appropriate search terms for a given dataset. The physicians could then manually set the context of each appearance of each term as having either a positive or negative context. Each document was then marked for relevance to a given query. To be marked *relevant*, a document needed to have at least one appearance of a query term tagged as a positive.

The following sentences demonstrate four examples taken from the training set which contain the term *nausea* in different contexts. The context of the term nausea in sentences 1 and sentence 2 was marked as positive, whereas in sentence 3 and sentence 4 the context was marked as negative.

- She had slight nausea, which was controlled with Zofran.
- The patient presented with episodes of nausea and vomiting associated with epigastric pain for the past 2 weeks.
- The patient was able to tolerate food without nausea or vomiting.
- The patient denied any nausea or vomiting.

Table 2: Context Distribution in the Training Set

Term	Positive Context (documents)	Positive Context (instances)	Negative Context (documents)	Negative Context (instances)
Nausea	284	370	251	286
Abdominal Pain	210	284	82	91
Weight Loss	94	108	21	21
Tobacco	95	97	110	113

Table 2 presents the distribution of negative and positive contexts in the training set. The data presented includes both the number of documents containing a given term and the number of total instances of a term's appearance. A given term can appear more than once in the same document.

Benchmark Algorithms

The study algorithm was compared to well-known, supervised induction methods: Decision Tree using the C4.5 algorithm [10], Naïve Bayes [11], Support Vector Machines using the improved Platt's SMO Algorithm [12], Neural Networks and Logistic Regression with a ridge estimator [13]. To use the above methods, a suitable procedure was developed for creating the dataset. Each instance in the dataset refers to a single sentence in the text that contains at least one of the investigated medical terms. All input attributes are Boolean, indicating whether a certain token (usually a word) has appeared in the sentence. The context of the sentence is referred to as the target attribute.

Additionally, the performance of context insensitive retrieval was measured; namely, assuming that the context is always positive. The last measurement is useful for determining the impact of context in retrieval from medical narratives.

Results

Table 3 presents the negative context profile obtained by the algorithm studied. This profile contains only ten words and/or phrases. Most of the entries in the table relate to the negative context. It is interesting to note that the term "no" and "not" are not included in this profile. Apparently, the mere presence of the word "no" or "not" is not sufficient to indicate negation.

Table 3: Profile Content for Negative Context

any	denies	of systems
change in	had no	was no
changes	negative for	without

Table 4 presents the mean F-Measure (over all queries) obtained by each query method on all medical terms. Note that the study algorithm obtained the highest F-Measure. Decision trees and support vector machines achieved the second best results.

Table 4: Benchmark Results for Various Query Methods

Method	Precision	Recall	F-Measure
Decision Tree	90.0	92.0	90.99
Support Vector Machines	93.9	87.5	90.59
Naïve Bayes	82.3	92.6	87.15
Logistic Regression	79.5	85.8	82.53
Neural Network	62.8	97.7	76.46
Context Insensitive Retrieval	100.00	75.51	60.65
Study Algorithm	99.57	95.45	97.47

Table 5: Comparison of Performance by Term

Query	Decision Tree			Study Algorithm		
	P	R	F	P	R	F
Nausea	95.83	95.83	95.83	100	97.98	98.98
Abdominal Pain	95.65	97.06	96.35	100	95.65	97.78
Weight Loss	88.24	100	93.75	100	91.18	95.38
Tobacco	89.19	91.67	90.41	97.37	92.5	94.87
Headache	92.15	95.35	93.72	100	96.32	98.13
Hypertension	83.02	93.63	88.01	100	97.73	98.85
Chills	87.62	98.34	92.67	96.66	94.54	95.56

For each query, the performance of the study algorithm is compared to decision trees, the best alternative algorithm per Table 4. Table 5 indicates that the study algorithm obtains better results in all queries and has relatively small variance. Furthermore, Table 5 reveals that the results obtained by the study algorithm for the previously unseen terms (*headache, hypertension* and *chills)* are similar to the results obtained in the remaining terms (*nausea*, *abdominal pain, weight* loss and *tobacco*).

Using McNemar's test with continuity correction [14], results from the decision tree classification are compared to results from the study algorithm. The Chi squared obtained is 11.172, with one degree of freedom. The two-tailed P value is 0.0008. By conventional criteria, this difference is considered to be statistically significant.

The Effect of UMLS Term Removal

Part of profile creation is filtering out words or phrases that appear in the UMLS Dictionary, thereby removing medical terms that are related to one another. As an example, examine negation detection for the term *nausea*, which is used in the training set. The term *vomit* may be included in the negative profile, since nausea is related to vomiting. However, when the resulting negative profile is used for querying other terms such as *hypertension*, the term *vomit* may be misleading. In this case it is not an indicator for negation. Terms that appear in UMLS usually are not good as generalized indicators.

The removal of related UMLS terms significantly improves the recall and, in some cases, improves the precision. On average, the removal of UMLS terms improves the F-measure by 8.28%. In addition, the profile length becomes 53% shorter. The impact on performance is shown in Table 6.

Table 6: The Effect of UMLS Terms Removal

Query	No Term Removal			UMLS Terms Removal			Improvement
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	
Nausea	96.59	85.86	90.91	100.00	97.98	98.98	8.07
Abdominal Pain	100.00	84.06	91.34	100.00	95.65	97.78	6.44
Weight Loss	100.00	73.53	84.75	100.00	91.18	95.38	10.64
Tobacco	96.77	75.00	84.51	97.37	92.50	94.87	10.36
Total	98.02	81.82	89.19	99.57	95.45	97.47	8.28

The Effect of Phrase Length

The influence of phrase length on the profile was examined. Sequences of consecutive one to five word phrases were tested. Profiles obtained indicate that no 4-word or 5-word phrases were found. Furthermore, all the 3-word phrases in the 3-word profile share a sub-phrase prefix that can be found in the 2-word profile. For example, the phrase "patient denies any," which is part of the 3-word profile, share the prefix phrase "denies," obtained in the 1-word profile. This is true for all phrases that contain more than three words. According to the retrieval algorithm, finding a phrase in the sentence is enough to define the context, hence finding an extended phrase (which contains the phrase as a sub-phrase) will not change the context designation.

Nevertheless, as shown in Table 7, searching for up to 2-word phrases significantly improves the precision for searching a single word. The F-measure significantly improves for the queries: "nausea" and "tobacco," slightly improves for "abdominal pain," and slightly worsens for "weight loss."

Table 7: The Effect of Phrase Length

Query	Single Word			Two Word Phrases		
	P	R	F	P	R	F
Nausea	76.74	97.98	86.84	100	97.98	98.98
Abdominal Pain	92	98.55	95.83	100	95.65	97.78
Weight Loss	94.44	100	97.14	100	91.18	95.38
Tobacco	69.81	92.5	79.57	97.37	92.5	94.87
Total	81.57	98.76	89.35	99.57	95.45	97.47

Error Analysis

Analysis of the cause for False-Positive and False-Negative results indicates that there are five main categories of error:

Compound sentence—Compound sentences are composed of two or more independent clauses, joined by a coordinating conjunction or a semicolon. For example, "There were no acute changes, but she did have a 50 pound weight loss." This sentence is built from two independent clauses connected by the word "but," which alters the context of the second clause. The proposed algorithm does not identify this alteration; therefore, a query for positive weight loss will fail due to the word "no" in the beginning of the sentence.

Future reference—In this case, the patient is given instructions on how to react to a symptom he may develop. For example, "The patient was given clear instructions to call for any worsening pain, fever, chills, bleeding." In this case the patient does not suffer from fever, chills or bleeding and a query for one of these symptoms will mistakenly retrieve the document.

Negation indicating existence—Although the meaning of a word might be negative, the context in which it is written might indicate otherwise. For example, "The patient could not tolerate the nausea and vomiting associated with Carboplatin."

Positive adjective—A sentence is written in a negative form, but an adjective prior to one of the medical terms actually indicates its existence. For example, "There were no fevers, headache or dizziness at home and no diffuse abdominal pain, fair appetite with significant weight loss." The adjectives "fair" and "significant" in the sentence indicates that the following symptoms actually do exist.

Wrong sentence boundaries—Sometimes the boundary of a sentence is not identified correctly. In this case, one sentence is broken into two, or two sentences are considered as one. For example, "She denies any shortness of breath, dyspnea, chest pain, G.I. bleed, fever or chills." In this case, the terms bleed, fever and chills were not associated with the negation, as the negation phrase was part of the first sentence. The retrieval algorithm did not detect the negation and mistakenly retrieved the document.

Figure 3 presents the distribution of errors in the test set. Note, compound sentences are responsible for majority of the errors.

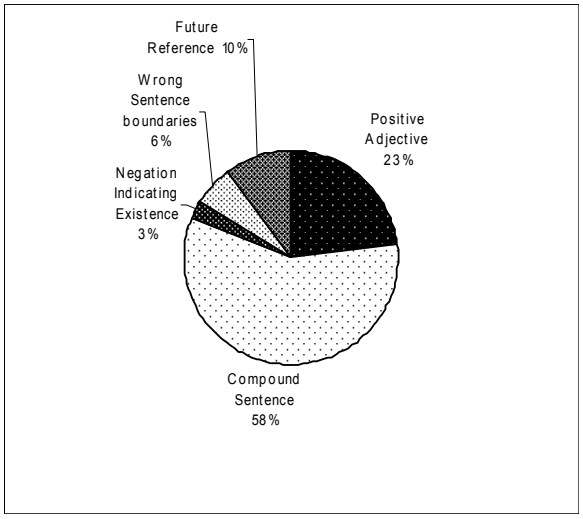

Figure 2 - Distribution of Errors

Conclusion

A new algorithm for identifying context in free-text medical narratives is presented. It has been shown that this algorithm is superior to traditional classification algorithms. The algorithm is not complex and is built from traditional building blocks that were adapted to the medical domain.

References

[1] Pratt A.W. Medicine, computers, and linguistics. *Advanced Biomedical Engineering* 1973: 3:97-140.

[2] Hersh W, Hickam D. A comparison of retrieval effectiveness for three methods of indexing medical literature. *American Journal of Medical Science* 1992: 303:292-300.

[3] Hersh WR, Hickam DH. Information retrieval in medicine: the SAPHIRE experience. Journal of the American Society of Information Science 1995: 46:743-7.

[4] Nadkarni PM. Information retrieval in medicine: overview and applications. *J Postgraduate Medicine* 2000: 46 (2).

[5] Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *J Am Med Inform Assoc* 2001: 8(6): 598-609.

[6] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanann BG. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics* 2001: 34: 301–310.

[7] Murff HJ, Forster AJ, Peterson JF, Fiskio JM, Heiman HL, Bates DW. Electronically Screening Discharge Summaries for Adverse Medical Events. *J Am Med Info Assoc* 2003: 10 (4): 339-350.

[8] Van Rijsbergen, CJ. *Information Retrieval*. 2nd edition, London, Butterworths, 1979.

[9] Lindbergh DAB. Humphreys BL. The Unified Medical Language System. In: van Bemmel JH and McCray AT, eds. 1993 *Yearbook of Medical Informatics*. IMIA, the Netherlands, 1993; pp. 41-51.

[10]Quinlan, J. R. C4.5: *Programs for Machine Learning*. Morgan Kaufmann, 1993.

[11]Richard Duda and Peter Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[12]S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murth, Improvements to Platt's SMO Algorithm for SVM *Classifier Design. Neural Computat* 2001: 13(3): 637-649.

[13]Cessie, S. and van Houwelingen, J.C. Ridge Estimators in Logistic Regression. *Appl Statistics* 1997: 41(1): 191-201.

[14]Dietterich, T., Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* 1998: 10: 1895-1924.

Address for correspondence

Tom H. Karson, M.D.
Chief Medical Information Officer
Continuum Health Partners
555 West 57th Street, 19th Floor
New York, New York 10019