International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 08 | Aug -2017 www.irjet.net p-ISSN: 2395-0072

LOCAL AND GLOBAL LEARNING METHOD FOR QUESTION ANSWERING APPROACH

RADHIKA S M[1], SYAMA R[2]

[1]SCT college of engineering, papanamcode, Trivandrum
[2]Assistant Professor, Dept. of computer science and engineering, Sct college of engineering, Papanamcode, Trivandrum

---

Abstract - *Vocabulary gap between health seekers and health care experts are more prevalent in health care domain. Different health seekers describe their questions in different ways and answers provided by the experts may contain non standardised terminologies. To overcome the vocabulary gap, a new scheme is used which combines two approaches namely local mining and global learning. Local mining extracts medical concepts from medical records and then map them to normalised terminologies based on standardized dictionary. Local mining suffer from problem of missing key concepts. Global learning overcome the issues in local mining by finding the missing key concepts.*

Key Words: **Local mining, Global learning, SNOMED-CT, UMLS.**

1.INTRODUCTION

Patients seeking online information about their health, connecting patients with doctors world wide to know about their health via question and answering. Doctors able to interact with many patients about particular issue and provides instant trusted answers for complex and sophisticated problems. Previously external dictionary is used to relate medical data which was not that much sufficient enough. Here we incorporate corpus aware terminology which is used to relate the natural language medical data with medical terminology this narrow down the path between health seekers and health providers. For example: heart attack can also be said as myocardial disorder. A tri-stage framework is used to accomplish the task.

_ Noun phase extraction
_ Medical concept detection
_ Medical concept normalization

Due to loss of information global learning approach is used to complement local mining approach. Medical sites are among the most popular internet sites today through which people can get more knowledge about their health conditions. The practice of medicine is experiencing a shift from patients who passively accept their doctors orders to patients who actively took online information to know briefly about their health because doctors are very busy with many patients and hence they cannot give brief description about their health issue to each and every patient. This is the reasons why health seekers normally use online medical sites. Most of the medical sites such as mayo clinic, Medscape are consumer oriented and provide their sound advice about general medical topics. The vocabulary used is readily comprehensive when health seekers search for more detailed information about a very specific topic. Due to tremendous number of records have been accumulated in their repositories and in most circumstances user may directly locate good answers by searching rather than waiting for experts to answer. However users with diverse background do not necessary share same vocabulary, the same question may be written in different native languages which is difficult for other health seekers to understand to bridge vocabulary gap corpus aware terminology is used.

2. RELATED WORKS

A. An Automated System for Conversion of Clinical Note into SNOMED Clinical Terminology

SNOMED-CT consists of many medical concepts and relationships. Identifies the medical concepts in the text. It is mainly used for medical data retreival. System mainly comprises of three modules namely Augmented lexicon, term compositor and negation detector. Augmented lexicon traces the words that appears in the text and identifies the concepts that are also in the SNOMED-CT. SNOMED-CT descriptions are then made into atomic words. UMLS Specialist lexicon performs the normalisation. Normalisation involves the removal of stop words. A token matching algorithm is used. It identifies the SNOMED-CT description in the text and also retreives the related descriptions from the data structure. Matching matrix is used to identify the sequences. Negation identification is to identify the negative terms. Number of negative terms are present in clinical text . SNOMED-CT also contains numerous negation terms. For each negative term in the text there is a mapping in the SNOMED-CT. Mapping is performed based on the SNOMED-CT concept id. To detect other negative concepts a simple rule based negation identifier is used. Identifies negation terms of the form negation phrase (SNOMED CT phrase)* (SNOMED CT phrase)* negation phrase [2] SNOMED-CT consists of many qualifier values. Qualifier modifies the medical concepts. Qualifier words are separated from augmented lexicon during concept matching.

B. Meeting Medical Terminology Needs The Ontology-Enhanced Medical Concept Mapper

Concept mapper maps the semantically related terms to the query. Helps user to access online medical information. System integrates AZ noun phraser ,Unified Medical Language System(UMLS), WordNet and Concept Space. AZ noun phraser extracts the noun phrases from free text. WordNet is a online accessible ontology and consists of set of synonyms. For example, in WordNet, "injection" has three senses: "injection as the forceful insertion of a substance under pressure (no synonyms). injection as any solutions that is injected (as into the skin) (synonym: injectant). and injection as the act of putting a liquid into the body by means of a syringe (synonym: shot."[3]. The UMLS involves the Metathesaurus, the Semantic Net, the SPECIALIST Lexicon, and the online Knowledge Sources. Metathesaurus is used for synonyms. SPECIALIST Lexicon is integrated with AZ noun phraser. DSP algorithm is used with Semantic Net. Concept Mapper provides users with synonyms for the query. System consists of three phases. AZ noun phraser extracts the medical concepts from the query. In the second phase synonyms are obtained from WordNet and the Metathesaurus based on the query. In the third phase related terms are obtained based on Concept Space and the Semantic Net.

C. Medical coding classification by leveraging inter-code relationships

Medical coding converts the information in the patient medical records into codes. ICD-9 is used to code medical records. Code is assigned to a patient when a patient is provided with service and also after discharge the patient is provided with a code. Multilabel large margin classifier is used to study about the code structure.

D. Fast tagging of medical terms in legal text

Medical terms occurs in news, medical, legal text etc. this involves a method that tags the medical terms and finds the longest set of words with the medical terms. These set of words are then converted into medical term hash keys. Then finds the concept id associated with the hash keys. This system uses a probabilistic term classifier. A probabilistic classifier is a classifier that is able to predict, given a sample input, a probability distribution over a set of classes, rather than only outputting the most likely class that the sample should belong to. Probabilistic classifiers provide classification with a degree of certainty.

E. A joint local-global approach for medical terminology assignment

In community generated health services like Health Tap, WebMD etc there is vocabulary gap between health seekers and experts due to the non standardized terminologies used by the experts in their answers. Joint local and global learning approach is used to label question answer pairs. Local mining labels question answer pair by extracting medical concepts. Local mining suffer from missing key concepts. Global learning enhances local mining.

F. Exploiting medical hierarchies for concept-based information retrieval

Keyword based techniques sometimes does not identifies the medical terms. Concept based method overcomes the problem of keyword based methods. In this the text documents are turned into concepts as arranged in SNOMED-CT. 'Bag of concepts' representation of documents is used. Terms are converted to concepts using natural language processing tool.

G. Domain-Specific term extraction and its application in text classification

Domain specific terms are identified using the statistical method. Entropy impurity is used to measure the word distribution in the domain. Normalisation process is added to identify the more specific domain terms. Samples are drawn from N samples to decide the impurity. If impurity is a lower value then samples are of one category. If impurity gets maximum value then N categories are equally in all samples. Popular measurement for impurity measurement is entropy impurity.

H. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries

Narrative medical reports contain negative terms. A algorithm NegEX is used to identify the negative terms. NegEx consists of many phrases that filters the sentences that contain negative terms. NegEx also controls the negation phrases. Indexed clinical findings and diseases are given as input to the algorithm. Output is the negated phrase in the findings. First the algorithm preprocesses the sentences. Then the relevant phrases are indexed and the applies the negation algorithm to it. Atlast compares the negation phrases detected by the algorithm with the negations identified by the physicians.

2. METHODOLOGY

A. LOCAL MINING

Local Mining involves a three stage framework. First stage is the noun phrase extraction in which the noun phrase are extracted. In the second stage the medical concepts are detected using concept entropy impurity(CEI). CEI also measures the specificity of that concept in the particular domain. Finally normalisation is performed. Normalizes the medical concepts based on the authenticated vocabulary.

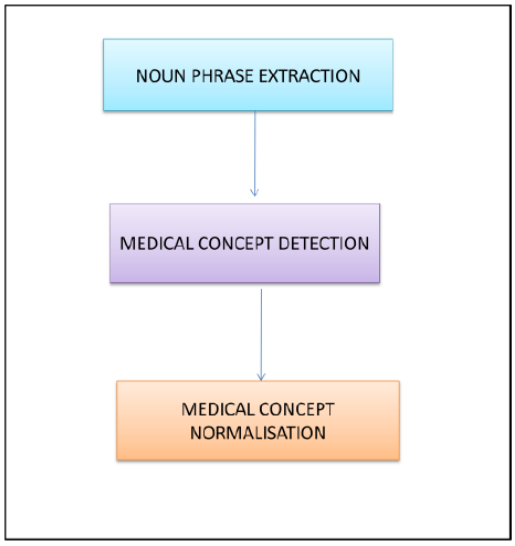

Fig -1: Local mining

1) NOUN PHRASE EXTRACTION:

Natural Language Processing is an upcoming field in the area of text mining. As text is an unstructured source of information, to make it a suitable input to an automatic method of information extraction it is usually transformed into a structured format. Part of Speech Tagging is one of the preprocessing steps which perform semantic analysis by assigning one of the parts of speech to the given word. Part-of-speech tagging (POS tagging or PoS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text. Part-of-speech tagging (tagging for short) is the process of assigning a part-of speech marker to each word in an input text. Because tags are generally also applied to punctuation, tokenization is usually performed before, or as part of, the tagging process: separating commas, quotation marks, etc., from words and disambiguating end-of-sentence punctuation (period, question mark, etc.) from part-of-word punctuation (such as in abbreviations like e.g. and etc.).In noun phase extraction, it takes the speech types part into account. In this process many unwanted words are stopped because that words are uninterested meaning. To extract the noun phrases, speech tags are assigned by Stanford POS tagger to every word of medical record given by the user. Then pulls out the sequence of words that match with the fixed pattern. The input to a tagging algorithm is a string of words and a specified tagset. The output is a single best tag for each word. The noun phrases should contain zero or more adjectives or nouns, followed by an optional group of a noun and a preposition, followed all over again by zero or more adjectives or nouns, followed by a single noun. To make up a noun phrase, sequences of tags are matched in a pattern. While there are many lists of parts-of-speech, most modern language processing on English uses the 45-tag Penn Treebank tagset.

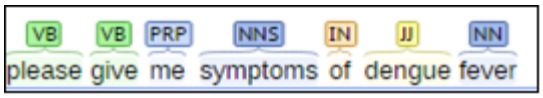

Fig -2: part of speech tagging

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	*+,%, &*
CD	cardinal number	*one, two*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb gerund	*eating*
JJ	adjective	*yellow*	VBN	verb past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, sing.	*IBM*	$	dollar sign	*$*
NNPS	proper noun, plural	*Carolinas*	#	pound sign	*#*
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... -- -
RP	particle	*up, off*			

Fig -3: penn tree bank Tagset

Parts-of-speech are generally represented by placing the tag after each word, delimited by a slash. For example, Take that Book.
VB DT NN (Tagged using Penn Tree Bank Tagset)

2) MEDICAL CONCEPT DETECTION:

Medical Concept Detection detects the medical concepts and differentiates it from other phrases. Concept Entropy Impurity is used to analyses the specificity of the medical concept. Larger CEI value indicates more relevant the concept in that domain.

3) MEDICAL CONCEPT NORMALISATION:

Medical concepts may not be standard terminologies so it is necessary to normalise the concepts based on a authenticated vocabulary. Consider birth control as an example it is not a standard terminology so it is necessary to map it to contraception. Authenticated vocabularies are ICD, SNOMED-CT, UMLS. SNOMED-CT provides the core general terminologies. Local mining suffer from incompleteness due to the missing key concepts. Second problem is the lower precision this is due to the irrelevant medical concepts in the records.

B. GLOBAL LEARNING

An enhanced and novel approach of global learning is being built for enhancing the result of local coding.

1) RELATIONSHIP IDENTIFICATION: Inter-terminology and Inter-expert relationships are analysed from the medical records.

2) INTER-TERMINOLOGY RELATIONSHIP: Terminologies in SNOMED-CT are arranged in hierarchies. For example, viral pneumonia is-an infectious pneumonia is-pneumonia is-a lung disease. Terminologies may also have multiple parents. For example, infectious pneumonia is also a child of infectious disease[1]. This hierarchial representation improves the coding.

3) INTER-EXPERT RELATIONSHIP: Analyses the historical data of experts and checks whether the experts are in the same or related area. Jaccard coefficient is used to analyse the experts relationship .

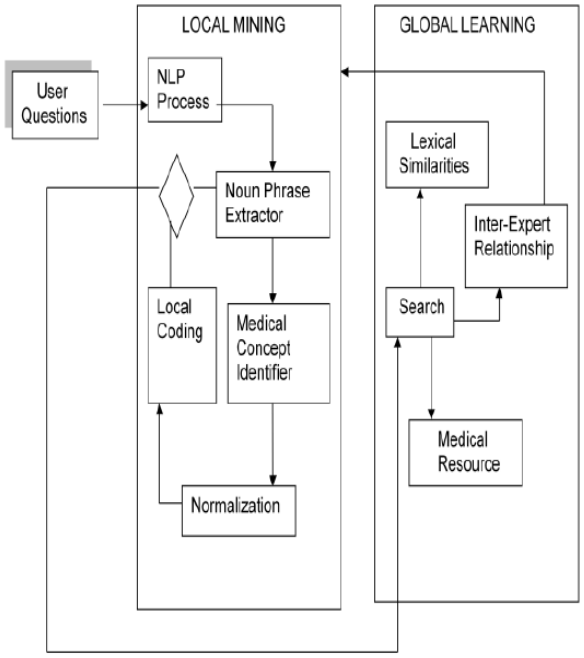

Fig -4: medical terminology assignment scheme

3. CONCLUSION

The proposed approach consist of a combined approach within the local mining and global learning, where the corpus aware terminology is being used for making a communication between the medical support seeker and the medical care providers. The corpus terminology is having the combined approaches of local mining and global learning, where the approach of local mining undergoes within the process of stemming, noun phrase extraction, spell check, normalization and detection of medical concept. The global learning maps the query against the indexed document or keyword that is relevant to the medical records. The query is being mapped within the local database and health seekers. The output is being produced based on the patients query.

REFERENCES

[1] Liqiang Nie, Yi-Liang Zhao, Mohammad Akbari, Jialie Shen, and Tat-Seng Chua, Bridging the Vocabulary Gap between Health Seekers and Healthcare Knowledge, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 2, FEBRUARY 2015.
[2] G. Leroy and H. Chen, Meeting medical terminology needs-the ontologyenhanced medical concept mapper, IEEE Trans. Inf.Technol. Biomed.vol. 5, no. 4, pp. 261270, Dec. 2001.
[3] Y. Yan, G. Fung, J. G. Dy, and R. Rosales, Medical coding classification by leveraging inter-code relationships, in Proc. ACMSIGKDD Int. Conf. Knowl. Discov. Data Mining, 2012, pp.193202.
[4] S. V. Pakhomov, J. D. Buntrock, and C. G. Chute, Automating theassignment of diagnosis codes to patient encounters using example-based and machine learning techniques, J. Amer. Med. Inf.Assoc., vol. 13, no. 5, pp. 516525, 2006.
[5] C. Dozier, R. Kondadadi, K. Al-Kofahi, M. Chaudhary, and X. Guo, Fast tagging of medical terms in legal text, in Proc. Int. Conf. Artif. Intell.Law, 2007, pp. 253260.
[6] M.-Y. Kim and R. Goebel, Detection and normalization of medical terms using domain-specific term frequency and adaptive ranking, in Proc. IEEE Int. Conf. Inf. Technol. Appl. Biomed., 2010, pp. 15.
[7] S. Hina, E. Atwell, and O. Johnson, Semantic tagging of medical narratives with top level concepts from SNOMED CT healthcare data standard, Int. J. Intell. Comput. Res., vol. 2, pp. 204210, 2010.
[8] H. Stenzhorn, E. Pacheco, P. Nohama, and S. Schulz, Automatic mapping of clinical documentation to SNOMED CT, Studies Health Technol. Inform., vol. 158, pp. 228232, 2009. . Intell. Comput. Res., vol. 2, pp. 204210, 2010.
[9] Y. Chen, Z. Chenqing, and K.-Y. Su, A joint model to identify and align bilingual named entities, Comput. Linguistics, vol. 39, no. 2, pp. 229266,2013.
[10] L. Nie, M. Akbari, T. Li, and T.-S. Chua, A joint local-global approach for medical terminology assignment, in Proc. Int. ACM SIGIR Conf.,2014.