Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

Using SNOMED CT Concepts for PAIRS

A.M. Mohan Rao, MB BS., MS,
Logic Medical Systems, Hyderabad, AP, India
ammohanrao@logicmedicalsystems.com

SNOMED CT medical vocabulary can be used to identify complementary features in a database. This functionality is used to develop a natural language processor (NLP) for PAIRS (Physician Assistant Artificial Intelligence Reference System). Although about 99% of concepts in PAIRS are present in SNOMED CT some features missing in it makes it unacceptable for any diagnostic decision support system (DDSS). Here we show that implementation of another NLP along with SNOMED CT makes it practically useful.

INTRODUCTION

Ideally medical databases must have clinical entities whose features comprise of medical domain , apart from microbiological, pathological, radiological and surgical domains. Computerization of such a data enables many interesting functionalities. Apart from being a learning tool it can also be a new source of knowledge which is of use in diagnosis. For example, feature-disease or feature-feature links can be deduced from disease-feature links. One can aim to achieve a uniform usage of clinical terms from such databases.

Utility of a database is limited by its completeness. Use of such a database helps one to design a Natural Language Processor which helps in data extraction from different data sources. Here, we show how SNOMED CT is a source of uniform clinical terms that not only can be used to simplify PAIRS database but also helps code a NLP that can be used for further evaluations. In this paper we use the terms PAIRS and SNOMED as representing their respective databases.

SNOMED vocabulary is used for several applications including Electronic Patient Records (EPR). However, as an evolving database many clinical terms are still missing in SNOMED which may be an impediment in developing a fully functional NLP. Here we show that this problem can be overcome by using a substitute algorithm (AINLP) (in addition to SNOMED algorithm) that works on PAIRS database.

PAIRS is an internet based DDSS that gives diagnosis for a given patient data. It is available free for evaluation from www.lmspairs.com upon request. Its artificial intelligence (AI) is based on a variational probabilistic belief networks as developed by Jaakkola & Jordan [1]. PAIRS database has 547 internal medicine diseases, 3700 unique features and 40 000 disease-feature links. Feature-disease links of a patient data are extracted from PAIRS database by AINLP. This process is limited by ability of AINLP algorithm to find a complementary feature in the database [2]. This limitation is rectified by SNOMED CT algorithm that works on SNOMED CT database.

There are several technological difficulties involved in NLP which may be caused by the algorithm, its implementation or its run times. These aspects are discussed at the end. Here we show how SNOMED CT can be used as a NLP for a DDSS.

MATERIALS

The computations are performed using HP Pavilion Entertainment PC with (1.6 GHz Intel processor, 1014 MB RAM) MS Vista operating system. We used a Perl program eutils written by Oleg Khovayko of National Library of Medicine to download about 2.5 million abstracts from PUBMED. NCBI Clinical Queries Research Methodology Filters developed by Haynes RB. et al is used. Routine search and extraction processes are done by MS Visual C++ programming language. A customized database PAIRS-DB is used to store and extract data programmatically. We obtained an International affiliate license for SNOMED CT from National Library of Medicine, USA. Sun Micro-systems Net Beans 5.5.1. IDE along with Visual Web Pack is used for internet enabling PAIRS. PAIRS database is owned and developed by Logic Medical Systems (www.logicmedicalsystems.com).

PAIRS

PAIRS comprises about 75000 disease-feature links for over 1700 diseases. This database is developed over a decade from various text sources and journals. We used a perl program eutils (written by Oleg Khovayko of National Library of Medicine) to download up to 2500 abstracts for each of the 1700 diseases from PUBMED. Queries based on Research Methodology Filters developed by Haynes RB. et al., are used for searching PUBMED [3]. We programmatically searched for each of the features in these abstracts. Initially MS Access is used to store the data as disease-feature links. It has over 7000 unique features. About 540 diseases having substantial number of features (about 40 000) are identified for application of artificial intelligence (AI) for diagnosis.

METHODS

SNOMED CT indices

We used SNOMED algorithm for assigning indices for PAIRS features. Over 3100 features are used as input data for this analysis. The search is facilitated by using a customized database known as PAIRS-DB. PAIRS-DB is programmatically loaded with SNOMED CT indices of duel_keywords, keyword and index-relationships. Each of the feature's duel_keyword and keyword indices are found using SNOMED algorithm. The indices (base-indices) are looked for inter-relationships in SNOMED CT relationships table. Those indices (relative-indices) having maximum frequency are selected as representative of the clinical feature. If a feature does not have any relative-index its base-indices are used for searching a concept. We used 250 clinico-pathological cases (CPC) published in New England Journal of Medicine (NEJM) between 1996-2003 for studying SNOMED CT based NLP functionality.

AINLP

AINLP is a substitute NLP that works independently of SNOMED algorithm. Its component database tables include: (a). general words: valid medical word and word pieces, (b).abbreviations: meanings of abbreviations. (c). synonyms, (d). antonyms and (e). feature-count: number of feature-disease links for given feature in PAIRS database. Input features from a patient data file are separated into their constituent words and their derivative word pieces. Word pieces are derived by deleting single letters from the end of each word in a cyclical fashion. These are checked against "general word list" for selecting only valid medical words. Words are searched for abbreviations and if found their meanings are selected. Further, synonyms and their antonyms are searched for the given words. Finally, the input data is searched for their complementary features in PAIRS database.
Computational times for AINLP are far shorter than SNOMED CT based NLP by several degrees. For example, for a list of 10-15 input features AINLP can find complementary features in 1-3 seconds where as same for SNOMED CT involves much long (sometimes as much as 1-3 seconds for each feature). Hence, AINLP is always run and if no complementary feature is found then only SNOMED CT based NLP is used.

SNOMED CT vs AINLP

For a given input feature duel keywords, keywords, their indices and relationships between indices are generated using SNOMED CT algorithm as described in SNOMED CT Clinical Terms Technical Implementation Guide [4]. Tables used for index search are: sct_concepts_duelkeyindex_20070731, sct_concepts_wordkeyindex_20070731. PAIRS-DB has 27 folders alphabatically labelled (plus a base folder that takes numerical and non-alphabetic data). Each of the folder is again assigned 27 folders. To reduce the computational time both the tables are programmatically loaded into PAIRS-DB. Finally, sct_relationships table has only those indices that are represented by PAIRS features. SNOMED CT based NLP runs in the following way: (a). find duelkeyindices, (b). find keyword indices that share duelkey indices . Finally, find maximally represented indices in sct_relationships table which is a PAIRS complementary feature for the input feature.
For a given input feature AINLP converts it into its words, and word pieces (by deleting single letters in a loop from the end). General words or abbreviations in this pool are identified and a search of PAIRS list of features is made. Complementary features in PAIRS database are identified by finding those that represent maximally in a search. Since volume of information processed in AINLP is much smaller than SNOMED CT based NLP AINLP computational times are much shorter.

Evaluation of NLP

PAIRS NLP has two components: AINLP and SNOMED CT. We tested the functionality of each in a two stage process. Firstly, we tested each component's ability to identify complementary features in PAIRS database. Secondly, we verified its function by testing PAIRS diagnostic output. We used 250 CPC cases of NEJM for this study. Each of the case shares some features from 3100 unique features of PAIRS database. Since AINLP computational times are much shorter than those of SNOMED CT based NLP AINLP is used in PAIRS always. SNOMED CT based NLP is used only if no complementary feature is generated by AINLP.

RESULTS

Functionality of SNOMED CT indices

Initially we included 3100 features of PAIRS in our analysis and identified 31 concepts (1 out of 100) that are not represented in SNOMED CT. Table 3 gives the features in PAIRS database that are identified as missing concepts in SNOMED CT. Here, we show results of search of SNOMED CT concepts in PAIRS.
Multiple indices are assigned to a given concept in SNOMED CT (see table 1). This can be problematic

in assigning an appropriate index number to a complementary feature in PAIRS. This problem can be resolved by use of relationships table.Our analyses of finding SNOMED CT concepts in PAIRS, we look for common relationships between indices/concepts rather than indices/concepts themselves. A number that is generated at a higher frequency in our search is preferred in assigning it to a feature in PAIRS.

For example "acute abdomen" may imply "acute abdominal pain". Same index number (9991008) is shared between three different concepts. Index for "Acute abdomen" (920005) is also assigned to "Acute abdominal pain syndrome". If we check the relationships table, only common relationship between two different indices of "Acute abdominal pain" (9991008 & 116290004) is "Abdominal pain" (21522001). Given two different indices for "Abdominal pain" (9991008) and "Abdominal pain through to back" (74704000) the only common relationship could be one index number identified by concepts: "Abdomen" or "Abdominal structure" (113345001).

SNOMED CT can be used as a form of NLP by implementing its algorithm and assigning a unique identifier index for a complementary feature in a given database. Any abbreviation or concept in a database can be accessed by using this functionality.

For example, index numbers for the feature ESR is given in table 2. By assigning 416838001 index for the feature "Erythrocyte sedimentation rate raise" in PAIRS, a user can access correctly by either entering "ESR" or "Erythrocyte sedimentation rate".

SNOMED CT missing concepts

Several radiological terms used in routine clinical practice are missing in SNOMED CT database (Table 3). Some of these features occur in significant number of diseases making them indispensable for any DDSS. For example, on chest x-ray "tree in bud pattern", miliary infiltrates, (contrast) enhancing lesion, pulmonary nodules represent 18,19,12 and 45 diseases respectively in PAIRS database. "Focal neurological lesion" missing concept represents about 27 PAIRS diseases. "Hypopigmented macules" and "heliotrope rash eyelids" are examples of other important missing concepts.

Evaluation of SNOMED CT indices

We assigned a SNOMED CT index to each of 3100 complementary features in PAIRS. We selected features from each patient data of 250 CPCs (NEJM) and looked for them in PAIRS using SNOMED CT algorithm. We are able to find the feature if input feature is represented in SNOMED CT. However, if a feature is not in SNOMED concepts, we are unable to find it in our results. We overcome this issue by using AINLP along with SNOMED CT. By testing features from each of patient data in 250 CPCs (NEJM) we are able to generate representative features in PAIRS. These representative features constitute not only the exact input feature but also features related to it.

Functionality of NLP

PAIRS NLP has two components: AINLP and SNOMED CT whose implementation of the algorithms is described in methods. AINLP is best suited for 3100 features listed in PAIRS. Advantages of it include the computational times which are rapid in finding complementary features. However, for those features which are not part of 3100 features, AINLP is not tested. In case where AINLP fails to find a complementary feature SNOMED CT algorith is used, thus the limitations of AINLP are supplemented by SNOMED CT.

We test the functionality of NLP by 250 CPC cases of NEJM. Each of the case has some of the features of 3100 features listed in PAIRS. Each case has about 10-30 features. The computational times are related
to the number of features in a case. If the features are more the time NLP takes more time to complete.

Index No.	SNOMED CT concept
9209005	Acute abdomen
9209005	Acute abdomen, NOS
158499006	[D]Acute abdomen
163250006	O/E - acute abdomen
207221008	[D]Acute abdomen
207255006	[D]Acute abdomen
268942007	O/E - acute abdomen
9991008	Spasmodic abdominal pain
9991008	Acute abdominal pain
9991008	Colicky abdominal pain
9209005	Acute abdominal pain syndrome, NOS
9209005	Acute abdominal pain syndrome
83132003	Upper abdominal pain
74704000	Abdominal pain through to back
71850005	Abdominal pain worse on motion
54586004	Lower abdominal pain
21522001	Abdominal pain
21522001	AP - Abdominal pain
116290004	Acute abdominal pain
111985007	Chronic abdominal pain
102614006	General abdominal pain-symptom
102614006	Generalised abdominal pain
102613000	Localised abdominal pain

Table 1 Multiple index numbers representing same SNOMED CT concept. This complexity can be resolved by finding relationships between them in relationships table.

Index No.	SNOMED CT concept
103208001	Erythrocyte sedimentation rate
103208001	ESR - Erythrocyte sedimentation rate
104154005	Erythrocyte sedimentation rate, non-automated
104155006	Erythrocyte sedimentation rate, automated
142875000	Erythrocyte sedimentation rate
165464006	Erythrocyte sedimentation rate
165466008	Erythrocyte sedimentation rate
165467004	Erythrocyte sedimentation rate
165468009	Erythrocyte sedimentation rate
165468009	Erythrocyte sedimentation rate
365649001	Finding of erythrocyte sedimentation rate
365649001	Erythrocyte sedimentation rate
365649001	Erythrocyte sedimentation rate - finding
416103000	Elevated erythrocyte sedimentation rate
416560009	Erythrocyte sedimentation
416838001	Erythrocyte sedimentation rate measurement
416838001	ESR - Erythrocyte sedimentation rate
416838001	Erythrocyte sedimentation rate measurement

Table 2 Natural language processor functionality by assigning unique identifier of SNOMED CT. See text for explanation.

During the test, the maximal computational times are in range of 15-20 seconds and each input feature finds its complementary feature correctly. This satisfactory results do not rule out possibility of features that may not give complementary features both by AINLP and SNOMED CT. It needs further testing to identify such features and take necessary steps.

DISCUSSION

SNOMED CT as NLP

SNOMED CT concepts are useful in many applications including EPR. Its application in Diagnostic Decision Support Systems (DDSS) is limited by presence (or absence of) a concept. It can simplify querying in a clinical database [5]. Its clinical vocabulary can be used for computerized diagnosis and problem list [6]. Almost complete coverage (98.5%) of concepts in SNOMED CT is reported. Out of 5000 features, about 92.5% concepts are covered in SNOMED CT [7]. Here, we report about 99% concept coverage in SNOMED CT in present study.

Missing concept in SNOMED CT	PMID
Air under diaphragm	7509402 (4)
Ataxia, sensory	8036880 (2)
Ataxia, stance	14561428 (1)
Spontaneous bleeding	14979383 (4)
Bowel wall thickness	16632735 (3)
Cervical sounds	
Edema of face	16340761 (6)
Exercise intolerance/ Exertional intolerance/ Effort intolerance	16689370 (5)
Nephromegaly	12621244 (1)
Focal neurological signs	16499723 (27)
Pulmonary oligemia	15658055 (4)
Heliotrope rash eyelids	10770031 (4)
Hypopigmented macules	15884465 (9)
Immobile diaphragm	*(2)*
Infundibular pinching	*(4)*
Leucoerythroblastic changes	*(4)*
Miliary infiltrates	11555380 (19)
Pre syncope	16195623 (9)
Pseudo fractures	6147751 (4)
Pulmonary nodules	15875070 (45)
Enhancing lesion	15891158 (12)
Secondary achalasia	11176337 (3)
Toxic granulocytosis	
Transient erythema	
Unilateral tongue weakness	12490688 (1)
Vertebral tenderness	7895748 (5)
Chest x-ray Hyperinflated_lungs	16338298 (5)
Chest x-ray honey comb appearance	*(5)*
Chest x-ray tree in bud pattern	*(18)*
Chest x-ray space occupying lesion of lung	*(2)*
X-ray skull space occupying lesion of brain	*(9)*

Table 3 PAIRS concepts missing in SNOMED CT. PMID shows related abstract number in PUBMED. Number in parenthesis shows number of diseases the feature is present in PAIRS. Feature-disease links for each are: (1). acute appendicitis (2). sub acute axonal polyneuropathy. (3). Wernicke-Korsakoff syndrome.(4). acute leukemia. (5). Crohn disease. (6). aortic incompetence. (7). aortic arch syndrome. (8). cardiac failure and dilated cardiomyopathy. (9). polycystic kidney disease. (10). AIDS related lymphoma. (11). CREST syndrome.(12). mixed connective tissue diseases. (13). ataxia telangiectasia. (14). liver abscess. (15). Hodgkin lymphoma. (16).megaloblastic anemia due to folic acid deficienc. (17). breast cancer. (18). arrhythmia. (19). cholestasis jaundice. (20). actinomycosis. (21). anterior

cerebral artery syndrome. (22). gastric carcinoma. (23). gram -ve septicemia. (24). *alcoholism (25). cerebral ischemia. (26). ankylosing spondylitis. (27*). *amyloidosis. (28). bronchiectasis. (29). allergic angiitis. (30). lung cancer. (31). epilepsy*.

PAIRS is a DDSS which has a database on which an artificial intelligence (AI) system works [2]. It has over 43 000 disease-feature links which are quantified and work in AI to give a diagnosis. Its AI system is based on variational probabilistic belief networks. For this, we require a robust NLP that can filter a clinical feature given any input. SNOMED CT is an ideal data source for such an application. However, its application is limited by a number of features missing. Many of these features may belong to radiology domain, which is crucial for diagnosis. For example, pulmonary nodules as a feature may be present in as many as 45 diseases in PAIRS but is missing in SNOMED CT. This suggests that use of SNOMED CT alone is insufficient NLP for PAIRS.

AINLP as NLP

For features which are exact complements of PAIRS or abbreviations of them, AINLP gives good results. However, results not shown here suggest that on its own AINLP is insufficient as NLP for PAIRS. This prompted us to use both AINLP and SNOMED CT for PAIRS. It is expected that those deficiencies in SNOMED CT are supplemented by AINLP. SNOMED CT has vast database and hence takes considerable amounts of time (up to about 3 seconds for each feature alone), we run AINLP first and if necessary (i.e., if a feature is not found by AINLP) then SNOMED CT is run. Thus, we check their computational run times.

SNOMED CT missing concepts

Typically diagnosis of a case involves not only history and physical examination but also interpretation of radiological data apart from pathological and microbiological data. Terms such "honey comb appearance" to describe a set of disease patterns is common in clinical setting. It is preferable to have these included in SNOMED CT concepts. As reported in results (see Table 2) many such concepts missing in SNOMED CT makes it impossible to use for a DDSS. Hence, AINLP is used along with SNOMED CT which substitutes the missing functionality. It is sometimes possible that a feature may not have its complementary for AINLP. In such a case SNOMED CT is allowed to run. We are not yet come across a feature that has no complentary for both AINLP and SNOMED CT. Presumably such a thing can happen either as a bug in the program or PAIRS-DB.

PAIRS as a DDSS

PAIRS is an internet enabled and can be used for diagnosing difficult cases. Features entered in a text area are processed by AINLP and SNOMED CT to select complementary features in PAIRS. These features are further processed by an AI system to give probabilities of diagnoses. These diagnoses are based on Bayesian probabilities and depend on age, gender and geographic parameters. Effectiveness of PAIRS functionality is limited by several technical and user difficulties. Generally, users like to enter patient data in a free form rather than choose from a table and they expect NLP to recognize the complementary features in the database for any given feature. For example, "myalgia" may suggest "bodypain", "bodypains", "body pain" or "body pains". However, "body" and "pain" are common for many (upto 30 000) concepts in SNOMED CT and hence its runtime process may become unacceptable (over 300 seconds). This problem is solved in PAIRS by using AINLP. PAIRS NLP consisting of both AINLP and SNOMED CT algorithms are tested using over 3100 unique features. Both the algorithms are complex and hence may yeild unpredictable outcomes in rare cases. The AI of PAIRS involves convex analysis and gives its diagnoses using a complex process. Therefore, activity of NLP may or may not affect diagnostic ability of PAIRS. It may not affect adversely if for example, "abdominal pain" finds a feature "abdominal pain, upper". But it may be otherwise if a complementary feature is not found at all. Many of these difficulties are minimized by use of AINLP followed if a complementary feature is not found by SNOMED CT algorithm.

Key advantages of PAIRS as a DDSS include not only its ability to generate a differential but also suggest procedures and features to look for in the patient for a given diagnosis. Its diagnostic process includes age, gender and geographic criteria based on epidemiological data from NHS, NCHS and WHO. PAIRS judgment on a diagnosis is graded into 7 heirarchial levels (certainly, as far as evidence goes, probably, necessarily, presumably, possibly and impossibly) on basis of variational probability, age, gender, geographic data, precipitating cause, duration, pathogenesis and system/ nonsystemic involvement. Highest grade prediction for a selected disease is attained only if all criteria match to the real data. For example, tuberculosis as selected disease does not match a geography "United States of America" because this presumption does not support

epidemiological data of NCHS. Hence, PAIRS judment never be "certainly" for such a diagnosis. PAIRS ability to suggest features to look for in a patient for given diagnosis is also very useful in arriving at a possible diagnosis.

PAIRS has a free component PAIRS-LM (which covers 980 diseases of which 580 are common internal medicine diseases) that gives links to about 2.5 million abstracts of PUBMED in National Library of Medicine (NLM). Each disease has about 2500 abstracts categorized into diagnosis, features, genetics, treatment, complications, prevention, incidence, nationality etc. This information is useful in the process of arriving at a diagnosis.

Two of the gold standards suggested for any DDSS include procedures to be performed for a given case and ability to extract data from EPR and give an output [8]. PAIRS suggests procedures for a given case. It also has a facility to select a case from multiple cases and give a diagnostic/ procedural output. However, several of technical difficulties discussed by Berner [8] such as correctness of diagnosis, quality of differential, user acceptness and amount of use are critical issues that still remain. Several additional advantages of PAIRS make it a possibly useful tool for arriving at a diagnosis atleast in difficult cases.

Technical problems

Main difficulty while using SNOMED CT arises because it has multiple indices assigned to a given concept. However, from its relationships table one can derive parent and child relationship between various indices (see results). Sometimes the relationships table may not yeild a clear parent/child ontologies for a given concept [9]. In such a case one may have problem in assigning an index to the complementary feature in user database. Results shown here are similar to those reported by others[9-10].

Computational times involving AINLP and SNOMED CT vary depending on the input feature. Typically AINLP run-times are much shorter than those for SNOMED CT and is run only if AINLP cannot generate any complementary feature. These difficulties prompt suggestions for users of PAIRS to get familiarize with PAIRS list of features before evaluating its diagnostic functionality and preferably to limit to those in 3100 list of PAIRS features as input data.

Acknowledgments

I thank National Library of Medicine, USA for giving an International Affiliate license for SNOMED CT. I thank Oleg Khovayko of National Library of Medicine, USA for Perl program e-utils.

References

1. Jaakkola, TS and Jordan, MI. Variational methods and QMR-DT database. J. Artificial Intelligence. 1999:10,291-322.
2. Mohan Rao.AM. PAIRS:a diagnostic decision-support engine. BJHC & IM. 2004:30-2.
3. Wilczynski, NL. and Haynes, R.B. Developing Optimal Search Strategies for Detecting Clinically Sound Causation Studies in MEDLINE. AMIA Annu Symp Proc 2003: 719-723.
4. SNOMED Clinical Terms Technical Implementation Guide, The International Health Terminology Standards Development Organization. July 2007. International Release.
5. Lieberman, MI. The Use of SNOMED CT Simplifies Querying of a Clinical Data Warehouse.AMIA Annu Symp Proc. 2003; 2003: 910.
6. Wasserman, H., Wang, J. An Applied Evaluation of SNOMED CT as a Clinical Vocabulary for the Computerized Diagnosis and Problem List. AMIA Annu Symp Proc. 2003; 2003: 699–703.
7. Elkin, PL, Brown, SH et al. Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists. Mayo Clin Proc.2006; 741-748.
8. Berner, ES. J Am Med Inform Assoc.2003, 10, 608-610.
9. Bodenreider, O. Smith, B.Kumar, A and Burgun, A., KR-MED 2004,12-20.
10. Ceusters, W., Smith, B., Kumar A and Dhaen, C. Mistakes in Medical Ontologies: Where Do They Come From And How Can They Be Detected? Ontologies in Medicine:Proc Workshop on Medical Ontologies. 2003.