

Comparing the Effects of Two Semantic Terminology Models on Classification of Clinical Notes: A Study of Heart Murmur Findings

Guoqian Jiang, Ph.D. and Christopher G. Chute, M.D., Dr. P.H.

Division of Biomedical Informatics, Mayo Clinic College of Medicine, Rochester, MN
(mailto:Jiang.Guoqian@mayo.edu)

Abstract

Objectives: We compared the effects of two semantic terminology models on classification of clinical notes through a study in the domain of heart murmur findings. **Methods:** One schema was established from the existing SNOMED CT model (S-Model) and the other was from a template model (T-Model) which uses base concepts and non-hierarchical relationships to characterize the murmurs. A corpus of clinical notes (n=309) was collected and annotated using the two schemas. The annotations were coded for a decision tree classifier for text classification task. The standard information retrieval measures of precision, recall, f-score and accuracy and the paired t-test were used for evaluation. **Results:** The performance of S-Model was better than the original T-Model ($p < 0.05$ for recall and f-score). A revised T-Model by extending its structure and corresponding values performed better than S-Model ($p < 0.05$ for recall and accuracy). **Conclusion:** We discovered that content coverage is a more important factor than terminology model for classification; however a templatestyle facilitates content gap discovery and completion.

Introduction

While modern terminologies have advanced well beyond simple one-dimensional subsumption relationships through the introduction of composite expressions, there is an emerging convergence of approaches toward the use of a concept-based clinical terminology with an underlying formal semantic terminology model (STM) [1]. SNOMED CT, the most comprehensive clinically oriented medical terminology system, currently adopts a foundation based on a description logic (DL) model and the underlying DL-based structure to formally represent the meanings of concepts and the interrelationships between concepts [2-3]. The existing SNOMED CT model is mainly pre-coordination oriented, i.e. containing many pre-coordinated terms, and also supports post-coordination. For example, a compositional expression “[*hypophysectomy* (52699005)] + [*transfrontal approach* (65519007)]” could be used to describe a more specific clinical statement than that only using the term “hypophysectomy (52699005)”. For a specific domain, a template model having a semantic structure with a coherent class of terms can be used as a formal representation [4]. This kind of model is mainly post-coordination oriented and a list of atomic terms is organized within a semantic structure.

For example, the latest version of the International Classification of Nursing Practice (ICNP) uses a 7-Axis model to support the representation of nursing concepts and integrates the domain concepts of nursing in a manner suitable for computer processing [5]. One of the main goals of the semantic terminology models is to support capturing structured clinical information that is crucial for computer programs such as information retrieval systems and decision support tools [6]. Structured recording has the potential to improve information retrieval from a patient database in response to clinically relevant questions [1]. However, functional difference in retrieval performance has not been clearly demonstrated between these two different semantic terminology models.

In this study, we focus upon the specific domain of heart murmur findings. Two schemas were established from two different semantic terminology models for evaluation: one schema is extracted from the existing SNOMED CT model (S-Model) and the other is a template model (T-Model) extracted from a concept-dependent attributes model recently published by Green, et al [7]. The objectives of the study are to annotate the real clinical notes using the two schemas and to compare and evaluate the effects of two models on classification of the clinical notes.

Methods and Materials

Defining the annotation schemas

We defined two schemas for both S-Model and T-Model and represented the two schemas in Protégé (version 3.2 beta), which is an ontology editing environment and was developed by Stanford Medical Informatics [8].

For the S-Model, we established a schema by extracting concept trees from the existing sub-hierarchy of heart murmur findings in January 2006 version of SNOMED CT (see Fig. 1). One root concept is “Heart murmur (SCTID_88610006)” which includes 86 sub-concepts of pre-coordinated terms of heart murmur findings. The other root concept is “Anatomical concepts (SCTID_257728006)” which includes two parts relevant to our schema. One part is the concept “Cardiac internal structure (SCTID_277712000)” and its sup-concepts. The other part contains only those anatomical concepts appearing in our clinical notes corpus on the basis of a manual review. For all heart murmur concepts, two semantic attributes derive from SNOMED CT context model for

heart murmur findings that frame post-coordination. One is “procedure site” that represents the auscultation site of a heart murmur and the other is “finding site” that represents the potential etiological site of a heart murmur. The values of the former one were set as the

instances of “anatomical concepts (SCTID_257728006)” and the values of the latter one were set as the instances of “Cardiac internal structure (SCTID_277712000)”.

Fig. 1 Schema of SNOMED CT Model (S-Model) for heart murmur findings represented in Protégé

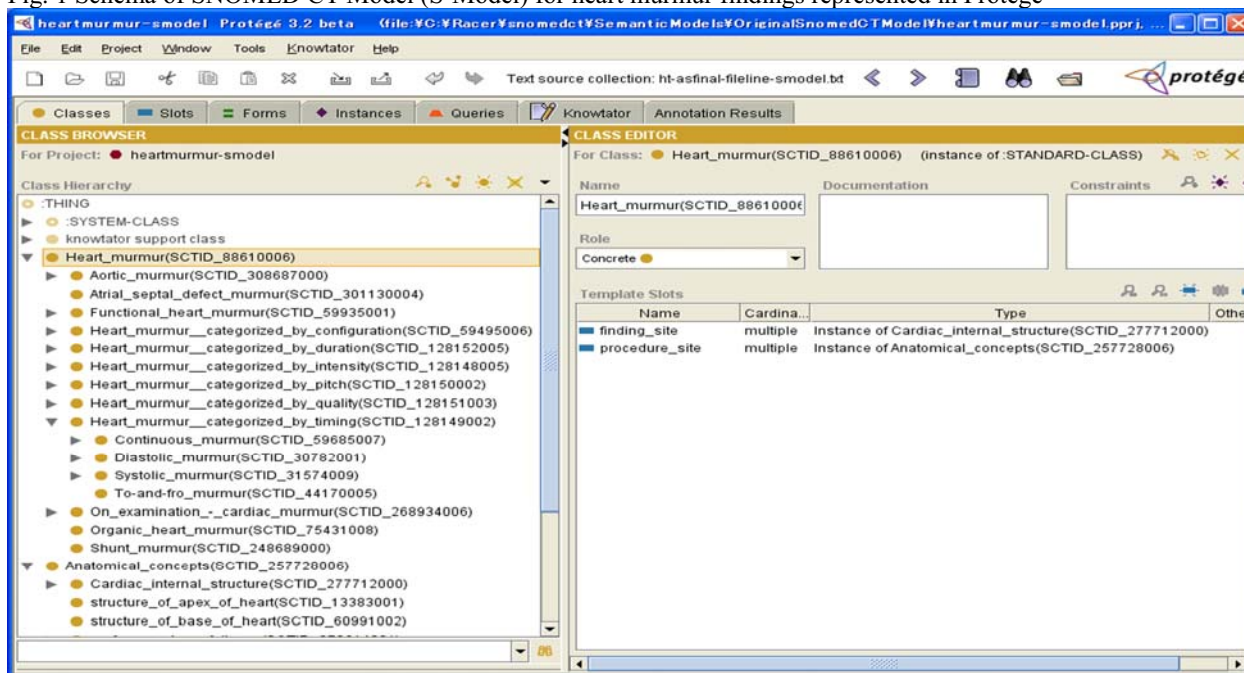
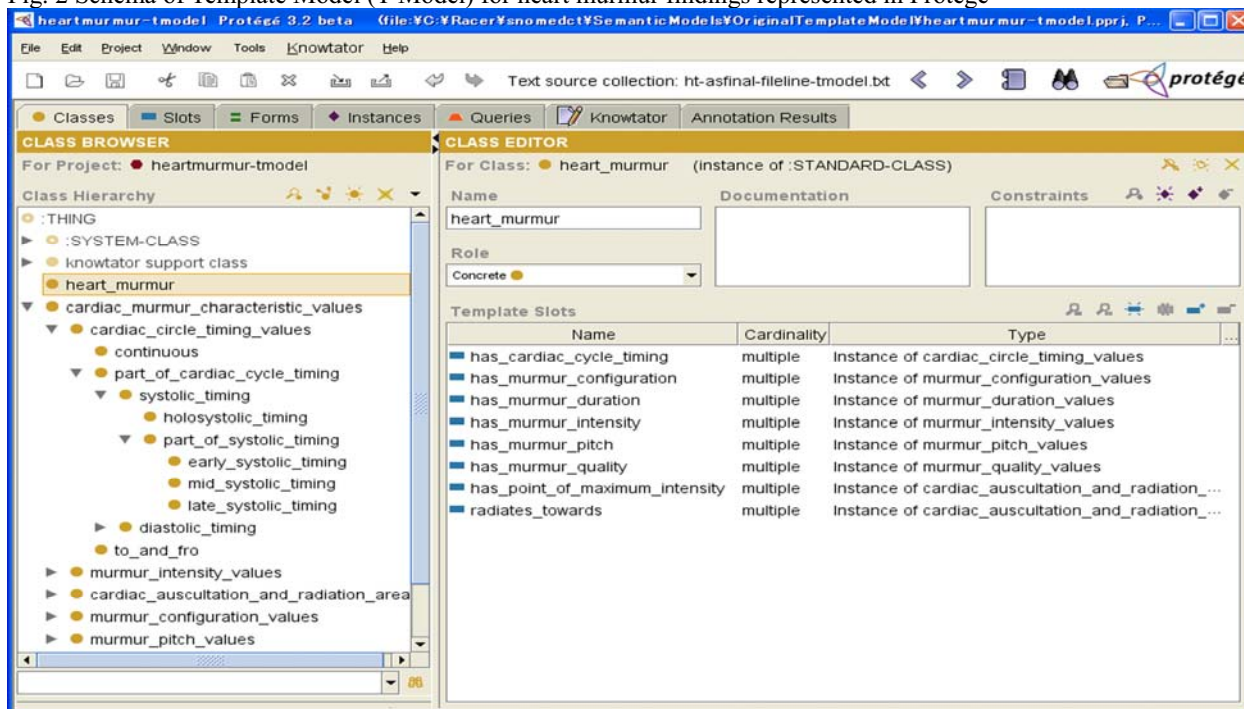


Fig. 2 Schema of Template Model (T-Model) for heart murmur findings represented in Protégé



For the T-Model, a schema was established from a concept-dependent attributes model published in a recent paper of Green, et al [7]. In this schema (see Fig. 2), one root concept is “heart murmur” which had eight semantic attributes, consisting of “has cardiac cycle timing”, “has murmur configuration”, “has murmur duration”, “has murmur intensity”, “has murmur pitch”, “has murmur quality”, “has point of maximum intensity”, “radiates towards”. The corresponding values of these eight attributes were set as the sub-concepts of the other root concept “cardiac murmur characteristic values”. We adopted the model attributes are directly from Green’s model, as well as their values (kindly provided by Green, interpersonal communication).

Preparing clinical notes corpus

The Mayo Clinic has a repository of approximately twenty million clinical notes that consist of documents dictated by physicians that are subsequently transcribed and filed as part of the patient’s electronic medical record. The following criteria were made to sample those notes. Firstly, we extracted notes with these criteria from Mayo repository in an automatic way: 1) created between January 1, 2005 to January 31, 2005; 2) Having a heart murmur description in *Physical Examination* section; 3) age ≥ 21 ; 4) Having a Hospital International Classification of Disease Adaptation (HICDA) code of the Heart Valvular Disease, and 5) removing patients with a code for status prosthetic valve or complication of a prosthetic valve. Secondly, we flagged extracted documents containing a diagnosis of aortic stenosis (AS), yielding 103 documents. Thirdly, we randomly selected controls among the extracted documents having no diagnosis of AS by matching the following conditions: 1) no history of valvular surgeries; 2) matching gender and age within 1 year for each case (see Table 1). Two controls were retained for each case, totaling to 309 documents. Finally, we parsed out cardiac exam from the *Physical Examination* section of each document to create an annotation corpus.

Table 1. Control documents selection by matching with gender and age

Age	Male	Control	Female	Control	Total
21-30	1	2	0	0	3
31-40	0	0	0	0	0
41-50	0	0	2	4	6
51-60	4	8	0	0	12
61-70	7	14	5	10	36
71-80	26	52	7	14	99
81-90	24	48	21	42	135
91-	2	4	4	8	18
Total	64	128	39	78	309

Annotation software and Annotators

A general purpose text annotation tool, Knowtator [9], was used to map text contents to our schema. Knowtator is a Java plug-in for Protégé and mainly used for creating gold-standard training and evaluation corpora for natural language processing (NLP) systems. The annotation schemas described in section above were instantiated in Knowtator.

One author (GJ) performed the annotation task and then the other author (CGC) verified the annotations for 10% of all documents. Differences were mutually adjudicated and lessons generalized to the remaining 90% of cases.

Coding for machine learning classification

We coded the annotated corpora for classification using a machine learning classification algorithm. The target category of the classification is binary, i.e. aortic stenosis (AS) or non-AS. In other words, the goal of the classification is to predict whether a document with a heart murmur description belongs to AS category or not. The annotations of each document were used as the predictive features and coded as binary.

We used a Weka implementation of the decision tree (J4.8) [10], which is a well-known supervised approach to classification.

Outcome measures and statistical analysis

For the annotation task, we compared the description completeness between the two models. The annotators were asked to judge whether the heart murmur descriptions of each document could be described completely through using the schema of a model while they performed annotation task. If they judged a document as “incomplete”, they indicated a reason for the judgment.

To evaluate the data retrieval task, we used the standard evaluation metrics of precision, recall, f-score and accuracy. Precision is defined as the ratio of correctly assigned AS category (true positive) to the total hit number (true positives and false positives). Recall is the ratio of correctly assigned AS category (true positive) to the number of target category in the test set (true positives and false negatives). The f-score represents the harmonic mean of precision and recall. Accuracy is the ratio of correctly assigned categories (true positives and true negatives) to total number of instances in test dataset.

For S-Model, one dataset (SM) that contains the annotations of both heart murmurs and anatomical concepts was prepared. For T-Model, three datasets were prepared. The first one (TM1) is that contains the annotations from Green’s original model. The other two datasets are extension of TM1. We extended TM1 to create TM2 by completing the values for all eight semantic attributes whenever a description appearing in the clinical notes corpus did not have a corresponding value in TM1. For example, we added “upper sternal border”, “mid sternal border” and “lower sternal

border” into the schema because they appeared frequently in our corpus to describe the auscultation areas and the original model only contains “sternal border”.

Building on TM2, we created our third model (TM3) by adding a new semantic attribute “has inferences to (specific murmurs or etiological mentions)” to the root concept “heart murmur” and also completing its corresponding values from those descriptions appearing in the corpus. We re-annotated all documents using the extended models respectively.

Ten-fold cross validation for retrieval was performed 10 separate times over all four datasets and the paired t-test was performed to test the statistical significance of performance measures between the dataset of S-Model and three datasets of T-Model.

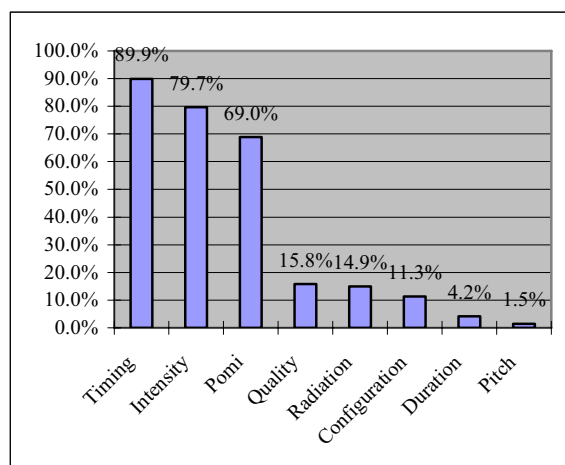
Results

For annotations

In S-Model, we made 995 annotations across all 309 documents. The average number of annotations per document is 3.2. Among the annotations, 728 belonged to 33 different sub-concepts of heart murmur (88610006). Of the heart murmur annotations, 509 (70.0%) had the values of the attribute “procedure site” filled and 6 (0.8%) had the values of the attribute “finding site” filled.

In T-Model, we made 1377 annotations against the original T-Model (TM1). The average number of annotations per documents is 4.5. Among 335 discrete heart murmur annotations, 89.9% include timing, 79.7% include intensity and 69.0% include points of maximum intensity (POMI). (see Fig.3)

Fig. 3 The annotation distribution of the eight attributes for all 335 heart murmurs annotated in original T-Model.



For comparison, the average number of annotations per document in S-Model was less than those in T-Model, indicating that S-Model supports more abstract way for description of heart murmur findings than T-Model.

Considering description completeness, 88 documents (28%) in S-Model were judged as “incomplete”; in the original T-Model, 201 documents (65%) were judged as “incomplete”. Thus, S-Model exhibits more complete domain coverage than the original T-Model.

The reasons for the incompleteness of four datasets from two models were listed in Table 2. We found that S-Model (SM) could describe most of “auscultation area” and the original T-Model (TM1) could not. For “radiation”, both SM and TM1 could not describe it well (we noticed that for SM, it is due to lacking of semantic attribute for “Radiation”, whereas that in TM1 is due to lacking of appropriate values for “Radiation” attribute). In addition, SM could describe all “ejection murmur” mentions and part of “aortic valve related” etiological mentions; TM1 could not. The results indicated that the strict template model, per Green, assumes that observers are using strict descriptions, and not making inferences to specific murmurs and etiological mentions, whereas SNOMED CT model accommodates partly the variability in inferences and strict descriptions, by providing terms that covers both.

Table 2 Frequency of reasons for the incompleteness of four datasets from two models

	SM	TM1	TM2	TM3
Auscultation area	1	78	0	0
Radiation	47	47	0	0
Configuration	8	8	0	0
Quality	7	5	0	0
<i>Specific murmurs</i>				
Ejection murmur	0	107	107	0
Regurgitant murmur	3	3	3	0
Flow murmur	2	2	2	0
<i>Etiological mentions</i>				
Aortic valve related	19	25	25	0
Mitral valve related	4	4	4	0
Pulmonary valve related	1	1	1	0
Septal defect	1	1	1	0

For TM2 and TM3, zero values in Table 2 indicated our synthetic completion of the values of each corresponding attribute in T-Model. The description completeness of TM2 was corresponding up to 57.6%, and that of TM3 up to 100%. Table 3 provided the examples (a AS case vs. a Non-AS case) to show how annotations were taken for all four schemas from two models.

For classification

As described in above section, four datasets (SM, TM1, TM2 and TM3) from two models were formed for evaluation. The results of the evaluation metrics of the four datasets were shown in Table 4. We found that the classification performance of SM was better than TM1 (i.e. original Green's model), with statistical significance identified for recall and f-score ($p < 0.05$, paired t-test). We consider that the reason was probably that the TM1 did not contain a complete list of murmur characteristic values for many of its semantic attributes.

The performance of TM2 was better than TM1, but still

lesser than SM. The result indicates that the original T-Model using strict physical descriptions may not fully represent descriptions of heart murmur findings in clinical notes, negatively impacting functional performance.

The classification performance of TM3 was the significantly best among the datasets ($p < 0.05$, paired t-test vs. SM). The result provided further evidence that inferences to specific murmurs and etiological mentions were important part of descriptions of heart murmur findings in real clinical notes, influencing the functional performance of the terminology model in this specific domain.

Table 3 The examples (AS Case vs. Non-AS Case) of annotations using four schemas

	AS Case	Non-AS Case
Textual Note	Heart: Loud 3 to 4/6 systolic ejection murmur heard best at the right upper sternal border. Absent of S2.	Heart: Regular rate and rhythm with a 2/6 left upper sternal border systolic regurgitant murmur. P2 was slightly increased. There was an S4 but no S3. The apical impulse was not localizable.
SM Annotation	15157000:Cardiac murmur - intensity grade III (VI) procedure site: [117144008:upper parasternal region] laterality: [24028007:right] 25311008:Cardiac murmur - intensity grade IV (VI) procedure site: [117144008:upper parasternal region] laterality: [24028007:right] 77197001: Ejection murmur procedure site: [117144008:upper parasternal region] laterality: [24028007:right]	36680007:Cardiac murmur - intensity grade II (VI) procedure site: upper parasternal region laterality: [7771000:left] 31574009: Systolic murmur procedure site: [117144008:upper parasternal region] laterality: [7771000:left]
TM1 Annotation	Heart murmur: has cardiac cycle timing value: systolic timing has murmur intensity value: intensity grade III/VI has murmur intensity value: intensity grade IV/VI has point of maximum intensity: sternal border (laterality: right)	Heart murmur: has cardiac cycle timing value: systolic timing has murmur intensity value: intensity grade II/VI has point of maximum intensity: sternal border (laterality: left)
TM2 Annotation	Heart murmur: has cardiac cycle timing value: systolic timing has murmur intensity value: intensity grade III/VI has murmur intensity value: intensity grade IV/VI has point of maximum intensity: upper sternal border (laterality: right) has murmur quality value: loud	Heart murmur: has cardiac cycle timing value: systolic timing has murmur intensity value: intensity grade II/VI has point of maximum intensity: upper sternal border (laterality: left)
TM3 Annotation	Heart murmur: has cardiac cycle timing value: systolic timing has murmur intensity value: intensity grade III/VI has murmur intensity value: intensity grade IV/VI has point of maximum intensity: upper sternal border (laterality: right) has murmur quality value: loud has inferences to: ejection murmur	Heart murmur: has cardiac cycle timing value: systolic timing has murmur intensity value: intensity grade II/VI has point of maximum intensity: upper sternal border (laterality: left) has inferences to: regurgitant murmur

Table 4 The results of the evaluation metrics of the four datasets

	Precision (mean±sd)	Recall (mean±sd)	F-score (mean±sd)	Accuracy (mean±sd)
SM	74.2% ±13.7%	59.4% ±15.6%	64.5% ±12.7%	79.0% ±6.1%
TM1	67.5% ±14.9%	*44.6% ±13.8%	*52.1% ±11.5%	73.6% ±5.4%
TM2	71.0% ±14.0%	53.2% ±18.9%	59.0% ±15.3%	76.9% ±6.8%
TM3	80.0% ±12.2%	*69.8% ±14.6%	73.5% ±10.4%	*83.6% ±5.8%

*p< 0.05 (paired t-test)

Discussions

In this study, we developed an approach to compare and evaluate the domain coverage (indicated by the description completeness) of two semantic terminology models and their effects on the classification of real clinical notes. We found that the description completeness of the S-Model was better than the original T-Model with original value set, correspondingly the performance of the S-Model on classification was also better. The extensions of T-Model that improved the description completeness, did improve its performance on classification of clinical notes. We clearly demonstrated that the domain coverage of a terminology model was directly correlated with its performance on classification of clinical notes; this is not surprising.

We could see that the effect of a terminology model on its functional performance in a specific domain mainly depends on its ability to represent the contents of the domain. In other words, the key issue for a terminology model is how to achieve complete domain coverage. If two different terminology models could represent the contents of a domain to achieve the same coverage, their performances on classification of clinical notes should have no difference.

In original T-Model, the description of a hear murmur could be fully post-coordinated by a semantic structure of eight semantic attributes. With original value set, we found that its description completeness was sub-optimal. In the paper from which the model was derived [7], the authors stated that “to adequately capture the full spectrum of cardiac murmur descriptions, our model needed a complete list of murmur characteristics”. So our first extension (TM2) completes the term values for all eight attributes of the original T-Model. The description of completeness was increased from 35.0% to 57.6%.

Thus, adding axes content to each attribute within the semantic structure did improve the domain coverage of the model; however, even with value completion, the original T-Model still could not achieve complete description for given corpus.

Therefore, we consider that the domain coverage of a

terminology model depends not only on the full value set of its semantic structure, but also on the coverage of the semantic structure itself.

Our second extension (TM3) of the T-Model adds a semantic attribute together with its corresponding values. This did overcome the limitation of semantic structure of the original T-Model and achieves a complete description for given corpus. In other words, the extended structure allows a systematic examination of where content gaps exist (e.g. missing values of references to specific murmurs and etiological mentions) and also guides the “completion” of the terms or missing contents informed by the extended structure.

In S-Model, most of its contents are pre-coordinated, with the post-coordination only possible for two semantic attributes “procedure site” and “finding site”. We did not extend the SNOMED CT model in a similar fashion since the model is an international standard although we believe that performance would be improved were it also extended. However, the extension of the model would be more complicated than that of template model because it involves both pre-coordination and post-coordination. We consider that the template model would be more applicable for achieving complete domain coverage. An important implication of these experiments is that a templatestyle terminology model more readily identifies gaps in coverage, and facilitates their completion for classification tasks.

Knowtator was used as our annotation tool and satisfied our purpose well, demonstrating the following merits. The first merit is that Knowtator uses the Protégé ontology editing environment to build the annotation schema. The frame-based knowledge representation system provides a flexible and expressive way to efficiently make schemas of the two model types in this study. The second merit is that Knowtator provides visualization of annotations, making the annotation task and confirmation process simple and efficient. The third merit is that the Java API of the system, which supports the annotation query that exports our coding of annotations to a classifier format automatically.

In order to improve the baseline performances on all standard evaluation measures, we performed control selection of clinical notes using strict criteria. This design did improve baseline performances (data not shown).

We regard the evaluation in this study in its comparative context across models; absolute measures of precision and recall are subject to factors beyond the scope of this study. A limitation of this study is that the annotations of clinical notes depends entirely on what clinicians decide to document for each patient, who they may or may not know has AS at the time. The local culture around documentation seems possible that these findings could be different on another corpus. Second, we only collected a relatively small size of clinical notes corpus given that the intensive annotation tasks were required. We consider that the annotation corpus is valid as both authors have clinical medicine background. Ten-fold cross validation used in this study may facilitate the efficient use of the data and get the best liability estimate. This kind of annotation corpus may be used to train a machine learning based annotation algorithm to build an automatic domain specific annotation tool. In addition, because it was not our intention to evaluate which classifier performed better, we only used a Weka implementation of the decision tree (J4.8) algorithm.

In conclusion, the domain coverage of the two models and their performance on classification clearly differ when applied to real clinical notes. Our approach provides an effective framework to evaluate the coverage and functional performance of the semantic terminology models in a specific domain for potential improvement. Future direction would focus on the scalability of the approach and the evaluation of interoperability among the different semantic terminology models.

Acknowledgements

This study is partly supported by NIH R01 LM07319.

The authors would like to thank Philip V. Ogren, Serguei V.S. Pakhomov, Guergana K. Savova, Pauline Funk and James D. Buntrock for their support.

References

- [1] Brown PJ, Sonksen P. Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model. *J Am Med Inform Assoc.* 2000 Jul-Aug;7(4):392-403.
- [2] Spackman KA, Campbell KE. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. *Proc AMIA Symp.* 1998;:740-4.
- [3] Yu AC. Methods in biomedical ontology. *J Biomed Inform.* 2006 Jun;39(3):252-66.
- [4] Zhou L, Tao Y, Cimino JJ, Chen ES, Liu H, Lussier YA, Hripcsak G, Friedman C. Terminology model discovery using natural language processing and visualization techniques. *J Biomed Inform.* 2006 Dec;39(6):626-36.
- [5] URL: <http://icn.ch/icnp.htm>; last visited at December 29, 2006.
- [6] Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. *J Am Med Inform Assoc.* 2006 May-Jun;13(3):277-88.
- [7] Green JM, Wilcke JR, Abbott J, Rees LP. Development and evaluation of methods for structured recording of heart murmur findings using SNOMED-CT post-coordination. *J Am Med Inform Assoc.* 2006 May-Jun;13(3):321-33. Epub 2006 Feb 24.
- [8] URL: <http://protege.stanford.edu/index.html>; last visited at December 29, 2006.
- [9] URL: <http://bionlp.sourceforge.net/Knowtator/>; last visited at December 29, 2006.
- [10] URL: <http://www.cs.waikato.ac.nz/ml/weka/>; last visited at December 29, 2006.