IJCLA VOL. 4, NO. 2, JUL-DEC 2013, PP. 81–99
RECEIVED 07/12/12 ACCEPTED 04/03/13 FINAL 14/05/13

SnoMedTagger: A Semantic Tagger for Medical Narratives

SAMAN HINA,[1,2] ERIC ATWELL,[1] AND OWEN JOHNSON[1]

[1] *University of Leeds, UK*
[2] *NED University of Engineering & Technology, Pakistan*

ABSTRACT

The identification and classification of semantic information in medical narratives is critical for various research applications such as question-answering systems, statistical analysis, etc. Our contribution is a novel semantic tagger named SnoMedTagger to tag complex semantic information (paraphrases of concepts, abbreviations of concepts, complex multiword concepts) with 16 SNOMED CT semantic categories in medical narratives. SnoMedTagger is developed to support domain users as well as non-domain users working on research questions using medical narratives. Our method includes corpus-based rule-patterns from real world dataset and rule-patterns developed by refinement of SNOMED CT (Systemised NOmenclature of MEDicine-Clinical Terms) clinical vocabulary. These rule-patterns were able to identify semantic information in a range of text and classify them with respective semantic categories derived from SNOMED CT. On unseen gold standard, our rule-pattern-based semantic tagger outperformed SVM-based machine learning system and Ontology-based Bioportal web annotator. The study has shown that it is possible to identify and classify complete semantic information with SNOMED CT semantic categories in medical narratives with high accuracy than achieved by existing approaches.

1 Background

The objective of this research was to develop a generic semantic tagger for identification and classification of semantic information in medical narratives. The presented tagger not only identifies complex multiword concepts, paraphrases of concepts and abbreviation of concepts but also provide a complete tagset extracted from SNOMED CT clinical vocabulary for classification of concepts. Researchers working in medical domain use different names for synonymous semantic categories for their specific research questions. For instance, semantic category 'Test' can also be referred to as 'Procedure' or semantic category 'Treatment' can also be named as 'Medications', which do not follow any standard names used in healthcare data standards. The SNOMED CT tagset used in this framework is customisable and can be used for classification of required semantic categories for various research applications using medical narratives. Because this tagset contains 16 semantic categories derived from international healthcare data standard SNOMED CT, therefore provide consistent information exchange among researchers with globally known semantic categories. SNOMED CT is globally the most comprehensive clinical vocabulary and is specified in several US standards (Stearns et al., 2001).

The classification of medical entities ('X-Ray', 'depression', 'No cough', etc.) with their semantic categories ('Procedures', 'Disorder', 'Findings', etc.) plays an important role in domain specific research. This semantic classification requires domain expertise which is time consuming and expensive; language researchers/non-domain researchers are dependent on domain experts to identify and/or annotate/classify domain specific information. In addition to this, it is also true that output of this approach, i.e., the annotated domain knowledge is restricted for specific research question(s) and therefore, cannot be reused by other researchers.

Many researchers developed biomedical named entity recognition taggers for classification of biomedical texts (Jonquet et al., 2009, Settles, 2005, Seth et al., 2004, Reeve and Han, 2007, Ananiadou et al., 2011). Some used SVM to identify and classify named entities in biomedical text (Zhenfei et al., 2011). Researchers mainly focused on the identification and classification of named entities using journal articles or MEDLINE abstracts but very few work is done on medical narratives with limited classification categories (Meystre et al., 2008).

Thus, there is a need to identify and classify not only named entities but complete semantic information in medical narratives. Medical narratives here refer to discharge summaries, progress notes, etc., written by clinicians whereas biomedical text refer to text in journal articles, MEDLINE abstracts, etc (Meystre et al., 2008). In medical narratives, clinicians express different concepts using semantics ('abbreviations', 'paraphrases', 'complex multi-word', etc.).

Researchers working on domain specific data have to spend considerable amount of resources in designing annotation guidelines and in hiring domain experts to identify and classify the required semantic categories in their dataset such as (Roberts A, 2007, Ohta et al., 2002, Wang, 2007). In automatic approaches, some researchers used linguistic patterns or ontologies to identify limited number of named entities in medical domain (Ogren et al., 2008, Mehdi Embarek and Ferret., 2008, Settles, 2005). Khare et al. (2012) performed contextual and structural analysis for mapping information on forms designed by clinicians with SNOMED CT concepts which is not suitable for unstructured information present in medical narratives.

Existing state-of-the-art systems such as Metamap and Bioportal provide ontologies for identification and classification of concepts in medical domain (Aronson, 2001, Noy et al., 2009a) and it has also been reported that Metamap does not perform well with medical narratives even with the use of extended modules (Meystre et al., 2008). In summary, the existing systems suffer from one or more limitations including failure at complex level of synonymy (Ogren et al., 2008), focus on any specific research question, corpus, limited number of semantic categories using controlled vocabularies/ontologies.

The identification and classification of semantic information from ever increasing number of medical narratives in patient records is critical and challenging for several research applications such as statistical analysis, question-answering systems, negation detection, relationship extraction, etc. In particular, we do not focus on mapping concepts with SNOMED CT controlled vocabulary but use SNOMED CT to classify concepts with semantic categories derived from SNOMED CT. This identification and classification will provide a consistent information exchange to domain users (medical/biomedical researchers) as well as non-domain users (language researchers).

As mentioned earlier, one of the major challenges is to cope with the informal writing structure which can vary from one clinician to another.

These variations in writing styles include the use of abbreviations, complex multi-word concepts, paraphrases of the concepts, etc (with/without use of punctuations). Thus, there is a need for a generic and comprehensive semantic tagger for medical narratives which should be flexible for a range of research questions and enables user to select semantic category according to their requirement. The present work describes the compilation of rule-pattern-based semantic tagger named SnoMedTagger by refinement of the international healthcare data standard SNOMED CT (version 2011) and analysis of real time dataset. Refinement of SNOMED CT was required because of limited writing structure of concepts in vocabulary. The evaluation proved that the SnoMedTagger is able to identify and classify concepts along with SNOMED CT semantic categories in medical narratives covering individual concepts as well as complete concept phrases.

In this paper, first we present the use of SNOMED CT semantic categories used in this research and the development of gold standard corpus for evaluation. Second, we describe the experimental setup of SnoMedTagger, SVM using uneven margins and existing Bioportal web annotator. Lastly, we present the evaluation of all systems against unseen gold standard test dataset and will discuss limitations and future directions.

2 Resources and Gold Standard Corpus

2.1 *Use of SNOMED CT*

In the present study, medical narratives were processed by Bioportal[1] 'Recommender' of ontologies and found the 'SNOMED CT' as best recommendation for medical narratives. SNOMED CT data standard was used for the following reasons; 1) The extraction of all concepts with their semantic categories from 'concept' table to develop a SNOMED CT dictionary application which was used to pre-annotate the corpus for the development of gold standard, 2) The refinement of 'SNOMED CT dictionaries' (explained in Section 3.1) which were used as base vocabulary and used in the development of rule-patterns for

[1] http://bioportal.bioontology.org/recommender

SnoMedTagger (SNOMED CT semantic tagger). Out of 31 top level concept classes and their sub-classes from SNOMED CT (Hina et al., 2010), concepts associated with 16 semantic categories (Attribute, Body Structure, Disorder, Environment, Findings, Observable Entity, Occupation, Person, Physical Object, Procedure, Product or Substance, Qualifier Value, Record Artifact, Regime/Therapy, Situation) were found in medical narratives used in this research. The remaining 15 semantic categories were missed due to following reasons;

- The semantic categories such as 'Physical force', 'Religion', 'Lifestyle', 'Staging and scales', etc were not found in the corpus used in this research. The concepts associated with these categories refer to special cases which can hardly exist in medical narratives.
- The concepts associated with the semantic categories such as 'Administrative concept', 'Link assertion' (For example; Has problem name, Has problem member etc), 'Namespace concept' (For example; Extension Namespace (1000145)), 'Inactive concept' (consists of outdated concepts, ambiguous concepts, etc), etc were to link and describe the other semantic categories in SNOMED CT data standard.

Particularly, we are not disambiguating the semantic categories in this research because some semantic categories ('Procedure – Regime/Therapy', 'Disorder – Findings') are closely related to each other. For instance, 'Regime/Therapy' is subclass of 'Procedure', 'Disorder' and 'Findings' are subclasses of 'Clinical findings' but according to domain experts may/may not be used as synonym in medical narratives and therefore should be classified separately. Also, it must be noted that semantic type named 'Product or Substance' is the combination of two separate top-level semantic categories, 'Pharmaceutical Product' and 'Substance' which were found synonymous in medical narratives.

2.2 *Development of the gold standard corpus*

The corpus used in this research was categorised into development dataset and test dataset. The development dataset was obtained from the fourth i2b2/VA 2010 challenge which contains discharge summaries and progress notes from different healthcare providers. The test dataset was provided by the Leeds Institute of Health Sciences. It consists of medical narratives written by medical students in a lab session in which

a consultation video was shown and the students recorded this consultation in 'System One', an EMR (Electronic Medical Record) system. Recorded narratives were then randomly extracted from the system to create an unseen test dataset. The medical narratives in test dataset were suitable to test the applicability of rule-patterns of semantic tagger as well as to evaluate the performance of the other two systems (SVM-based system, Bioportal web annotator).

The gold standard development dataset and test dataset were annotated following an instruction manual. The instruction manual was designed by authors, considering language issues identified in (Hina et al., 2011). This annotation scheme followed semi-automatic method which is feasible, cheaper and faster compared to manual annotation. This helped both types of users to complete the annotations on time. Two domain users annotated both datasets (development dataset and test dataset) independently following same annotation scheme. The inter-annotator agreement (IAA) was calculated between double annotated datasets as described by (Roberts A, 2007). The inter-annotator agreement for the gold standard development dataset and test dataset was very high and the disagreements were reviewed by a third domain expert. Test dataset was annotated in less time due to less number of concepts and achieved higher IAA than development dataset. Thus, the final gold standard for both datasets was compiled in a consensus set by adding disagreed concepts reviewed by third domain expert. Table 1 shows inter annotator agreement (IAA) and total number of SNOMED CT concepts in the final development and test dataset.

Table 1. Inter annotator agreement and number of annotated SNOMED CT concepts in gold standard development and test dataset

Gold Standard	IAA (%)	Concept annotations in final gold standard
Development dataset	86	5125
Test dataset	95.25	2672

3 Experimental Setup

This section includes the development of SnoMedTagger along with the implementation of the other two systems (SVM based supervised machine learning system, Bioportal web annotator) for evaluation.

3.1 *SnoMedTagger: SNOMED CT Semantic Tagger*

SnoMedTagger is a novel and comprehensive rule-pattern-based semantic tagger for the identification and classification of individual concepts, paraphrases of concepts, abbreviations of concepts and complex multiword concepts along with their SNOMED CT semantic categories in medical narratives. For the development of rule-patterns for semantic tagger, the dictionaries of 16 semantic categories were refined to develop rule-patterns for SnoMedTagger (explained in next section). Although rule-based approach require manual effort, still is effective in absence of large annotated corpus.

Refinement of SNOMED CT concepts for detecting individual concepts and abbreviations

For our purposes, we defined refinement as simplification of multiword concepts, separation of abbreviations from their definitions and removal of unnecessary concepts which are not used by clinicians. The dictionaries of semantic categories derived from SNOMED CT were refined in order to develop generic rule-patterns for SnoMedTagger. In following examples of refinement, all semantic categories are italicised while '→' represents refinement process.

Case 1: Removing unnecessary words and descriptions from SNOMED CT 'Concept' table

In SNOMED CT concept file, several multiword concepts contain descriptive information associated with them. Clinicians do not write this descriptive information in medical narratives and therefore it should be removed for accurate information extraction. Examples of removing descriptions such as 'NOS ', '[SO]', 'NEC', (structure) are as follows.

Example 1:
SNOMED CT concept: Skin NOS – *Body Structure*
Here, NOS = Not otherwise specified
Skin NOS – *Body Structure* → Skin – *Body Structure*

Example 2:
SNOMED CT concept: Vitreous membrane (structure) – *Body Structure*
Vitreous membrane (structure) – *Body Structure* → Vitreous membrane – *Body Structure*

Case 2: Simplification of multiword concepts into individual concepts

Multiword concepts were simplified into individual concepts to produce general rules for SnoMedTagger application following the steps shown below.

Example: SNOMED CT concept:
Entire Skin of Eyelid – *Body Structure*
Step 1: Entire Skin of Eyelid – *Body Structure* →
1) Entire Skin – *Body Structure*
2) Eyelid – *Body Structure*
Step 2: Entire Skin – *Body Structure* →
1) Entire – *Qualifier Value*
2) Skin – *Body Structure*

Case 3: Separation of abbreviations with their descriptions

Several studies reported the extraction of acronyms and abbreviations in biomedical text mainly MEDLINE abstracts using pattern-based approaches and regular expressions (Pustejovsky et al., 2001b, Pustejovsky et al., 2001a, Schwartz and Hearst, 2003). (Nadeau and Turney, 2005) adopted supervised machine learning approach for the identification of acronym-definition pair in biomedical text. (Ao and Takagi, 2005) proved corpus-based algorithm for the identification of abbreviations from MEDLINE abstracts.

In contrast, it was observed that clinicians prefer to write either short form (abbreviation) or long form (definition) in medical narratives. SNOMED CT contains abbreviations along with their definitions in the ontology and also stores this information separately which restrict writing styles in medical narratives.

For this reason, example case described here involves separation of abbreviations from their definitions for each respective dictionary. For instance, SNOMED CT concept: DVT – Deep venous thrombosis or DVT or Deep venous thrombosis can be written in other several possible forms; DVT – (Deep venous thrombosis), DVT (Deep venous thrombosis), (Deep venous thrombosis), DVT, Deep venous thrombosis, (Deep venous thrombosis) DVT, DVT (Deep venous thrombosis), (DVT), DVT: Deep venous thrombosis, Deep venous thrombosis: DVT.

Such concept and similarly other concepts containing abbreviation were simplified as follows:

DVT – Deep venous thrombosis – *Disorder* →
1) DVT – *Disorder*
2) Deep venous thrombosis – *Disorder*

However, there were no examples of abbreviation-definition pair in the development dataset, several pattern-based rules were developed to generalise SnoMedTagger on other datasets (medical narratives). The refinement of SNOMED CT dictionaries is an intermediate stage to apply generic rules for the extraction of semantic information from medical narratives.

System Flow of SnoMedTagger

SnoMedTagger application was developed using GATE - General Architecture for Text Engineering. GATE is an open-source natural language processing software which includes CREOLE: Collection of Reusable Objects for language engineering (Gaizauskas et al., 1996). CREOLE components were used to carry out basic language processing tasks (tokenisation, sentence splitting, part-of-speech (POS) tagging), morphological analysis, and gazetteers/dictionaries. Java Annotation Patterns Engine - JAPE transducers (Cunningham et al., 2000) were used to write rule-patterns for each SNOMED CT semantic category. SnoMedTagger application used 18 CREOLE components and 15 of them were based on JAPE transducers for the development of rules for 15 semantic categories (excluding 'Attribute'), as shown in Fig. 1.

The SnoMedTagger application pipeline first apply basic langauge processing resources (tokensiser, sentence splitter, part-of-speech tagger (Hepple, 2000)) on corpus.

Then, GATE processing resource called flexible gazetteer was used in SnoMedTagger pipeline for the detection of singular as well as plural concepts from refined SNOMED CT dictionaries (explained in earlier section). The flexible gazetteer provides the flexibility to customise the output of refined SNOMED CT dictionaries by morphological analysis. For detection of plural concepts, we used root feature of tokens.

After the identification of both singular and plural concepts with their respective semantic categories, set of rules were added in the SnoMedTagger. Semantic category 'Attribute' does not require rules; therefore rules were developed for the remaining 15 semantic categories.

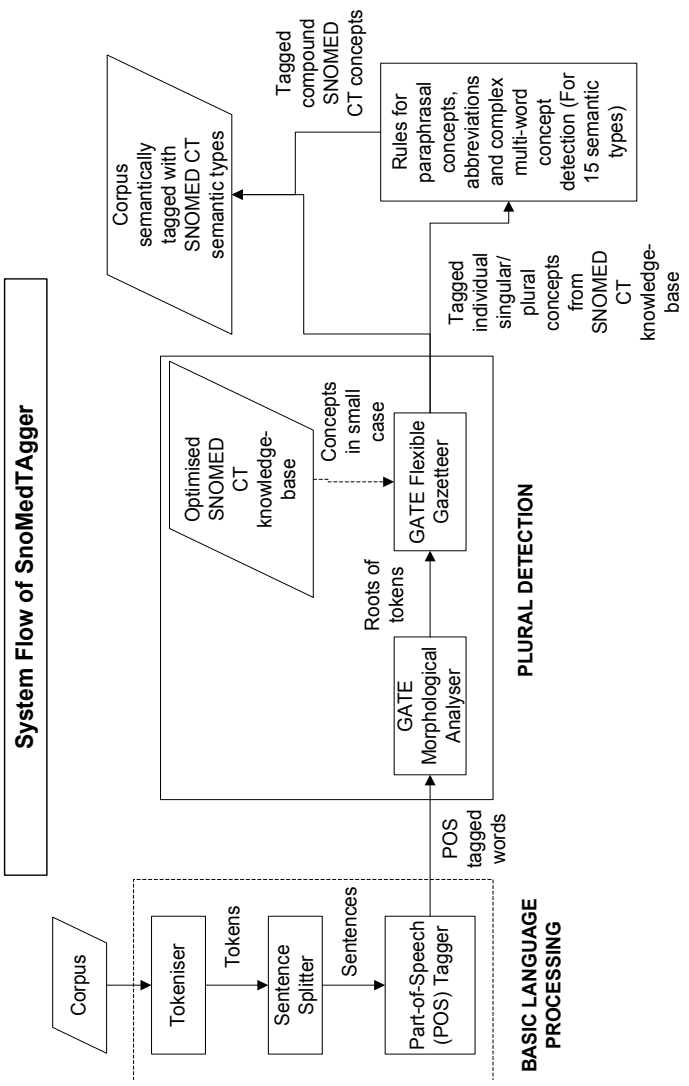


Fig. 1. Application pipeline of SnoMedTagger

This section will explain the development of rule-patterns for the identification and classification of paraphrase concepts, abbreviation of concepts and complex multiword concepts in medical narratives. The

derivation of quality rules was from two resources; 1) Analysis of the SNOMED CT data standard. 2) Language of medical narratives written by clinicians.

SNOMED CT data standard contains description logic which is meant to define ontology but has limitation of identifying concepts in medical narratives because of variation in writing styles. Therefore, the rule-patterns were written by analysing real world dataset (development dataset) and rule-patterns analysed during the refinement of SNOMED CT dictionaries. Rules-patterns were written as follows; Rule-pattern --> Rule-action. Example 1 show rule-patterns written by analysing language in SNOMED CT and example 2 contains rule-patterns written by analysing development dataset, where all the semantic categories are italicised. The other notations used in the examples are as follows:

sp= Space Token
IN= Preposition or sub coordinating conjunction
DT= Determiner
|=Or
Lookup.majorType = Bodystructure (dictionary of individual body structures such as 'chest', 'pelvis', 'leg', 'abdomen', etc.)
Lookup.majorType = Procedure (dictionary of individual procedures such as 'X-Ray', 'radiography', 'CT scan', 'biopsy', etc.)
Lookup.majorType = Qualifier_value (dictionary of individual qualifier values such as 'left', 'right', 'upper', 'lower', etc.)

Example 1:
SNOMED CT Concept:
'Radiography of chest' should be marked as *Procedure* and it can be written in several ways:

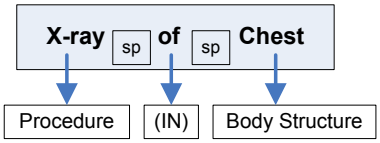
X-Ray of the chest
Chest X-Ray
Chest x-ray
Radiography of the chest
X-Ray of chest
X-ray of chest
CXR

The individual concepts such as radiography, X-ray, x-ray, X-Ray were marked by dictionaries/gazetteers while for the identification of multiword concepts and paraphrases, following rule-patterns were writ-

ten using dictionaries and linguistic features analysed in the development corpus.

```
Rule: Procedure
{
(Lookup.majorType = Procedure) (sp) (IN) (sp) (Lookup.majorType = Body structure) |
(Lookup.majorType = Procedure) (sp) (IN) (sp) (DT) (sp) (Lookup.majorType = Body
                                                                    structure) |
(Lookup.majorType = Body structure) (sp) (Lookup.majorType = Procedure)
}: label
-->
:label.Procedure = {Rule=Procedure}
```

For instance, first pattern in this rule can be described as follows:



These rule-patterns are general and will extract other concepts such as; 'GI Prophylaxis', 'pelvic lymphadenectomy', 'abdomen x-ray', 'Prostate biopsy','X-Ray of abdomen' and so on.

Example 2*:*
Below are some corpus-based rule-patterns analysed for the semantic category *Body structure*.

```
Rule: Bodystructure
{
(Lookup.majorType = Bodystructure) (sp) (IN) (sp) (Lookup.majorType = Body structure) |
(Lookup.majorType = Bodystructure) (sp) (IN) (sp) (DT) (sp) (Lookup.majorType = Body
                                                                    structure) |
(Lookup.majorType = Qualifiervalue) (sp) (Lookup.majorType = Bodystructure)
}:label
-->
:label.BodyStructure = {Rule=BodyStructure}
```

These general rule-patterns successfully identified concepts such as 'abdomen of the pelvis', 'Left leg', 'upper quadrant of the belly', 'left eye', 'chest wall', 'second toe on the right foot', 'left ventricular wall

thrombus', etc. Similarly, N=316 generic rule-patterns have been written for the 15 semantic categories by analysing all possible combinations of refined SNOMED CT dictionaries and linguistics features, shown in Table 2.

Table 2. Successful combinations of refined dictionaries and linguistic features used in the development of rule-patterns for SnoMedTagger. Shown are the 15 SNOMED CT semantic categories for which rules were developed.

Successful features used in the development of Rule-Patterns		Body Structure	Disorder	Environment	Findings	Observable Entity	Occupation	Organism	Person	Physical Object	Procedure	Product or	Qualifier Value	Record Artifact	Regime /Therapy	Situation
Token.features	Punctuation	■	■	■	■	■	■	■	▒	■	■		▒		▒	
	IN	■	■	■	■	■	■	■	▒	■	■			■	■	■
	DT	■	■	■	■	■	■	■	▒	■	■			■	■	■
	TO	■	■	■	■		■			■	■		▒	■	■	
	CC	■					■	■		■	■		▒			
	JJ		■													
	VBG			■							■					
	VBN			■												
Refined SNOMED CT semantic categories	Attribute		■		■	▒				■						
	Body Structure	■	■		■	▒				■	■		▒		▒	■
	Disorder		■	■					■		■			■	■	■
	Environment			■	■		■		▒	■	■		▒	■		
	Findings				■						■			■		■
	Observable Entity				■	▒										■
	Occupation			■	■		■				■		■	■		
	Organism		■					■								
	Person		■	■					▒		■		■	■	▒	
	Procedure		■			▒			■		■		■	■		■
	Physical Object			■	■		■		■	▒	■					
	Product or Substance		■		■	■	■				■	■			■	
	Qualifier Value	■	■	■	■				■		■					■
	Record Artifact												■	■		
	Regime /Therapy														■	
	Situation															■

LEGEND: Highlighted boxes indicate used features

3.2 *Using Supervised Machine Learning for Semantic annotation*

To evaluate the performance of our rule-based approach against machine learning, we used Java version of Support Vector Machines (SVMs) package LibSVM with uneven margins (Li and Shawe-Taylor, 2003). SVM is known for classification in language processing tasks and learns all features with high generalisation using kernel function. We used linear kernel with the extension of multiple classification ('one Vs others'). The general feature set used in the development of patterns was also used to train the classifier on development dataset (training set). The training was completed using following feature set.

1. Refined SNOMED CT dictionaries (for chunking individual concepts).
2. Part-of-speech categories of three words before and three words after dictionary terms.
3. Three Words before and three words after the roots of the token
4. The type/kind of tokens for learning punctuations 4 words before and 4 words after the term. These ranges were provided in order to learn long and complex multi-word concepts from the development corpus. The results were then compared against gold standard test dataset, described in section 4. Results showed that it is difficult to achieve high recall using general features for all 16 semantic categories.

3.3 *Bioportal Web Annotator*

Bioportal is a web portal which provides a selection of over 300 ontologies from biological and medical domain (Noy et al., 2009b). In this research, bioportal 'recommender[2]' was used for the recommendation of SNOMED CT ontology for medical narratives and then bioportal web annotator was used to annotate test dataset with selection of 16 SNOMED CT categories used in this research. Bioportal provide python client code which was used to annotate the test dataset using SNOMED CT ontology.[3] The annotations were then compared against human annotated gold standard presented in results section.

---

[2] http://bioportal.bioontology.org/recommender

[3] http://www.bioontology.org/wiki/index.php/Annotator_Web_service

4 Evaluation

The SnoMedTagger was developed using development dataset that contained concepts associated with 16 semantic categories derived from SNOMED CT; however the 'Organism' semantic category was missing in the gold standard test dataset. To evaluate all the three systems against unseen gold standard test dataset that contained 15 semantic categories, standard metrics (recall, precision, f-measure) were used. We focused on improvement of recall and f-measure of the SnoMedTagger to prove reliability of the rule-patterns. SnoMedTagger overall achieved 82% recall, 71% precision and 76% of f-measure while SVM based system overall achieved 49% recall, 81% precision, 61% f-measure and Bioportal system achieved 52% recall, 40% precision, 45% f-measure. The f-measure of rule-pattern-based SnoMedTagger outperformed the application using SVM with uneven margins (SVM-UM) and the ontology-based Bioportal web annotator. The application using SVM with uneven margins has achieved high precision but achieved very low recall because of granularity levels (identification of concept phrases).

On the other hand, ontology-based Bioportal web annotator predictably achieved low scores in all three systems because of inappropriateness of controlled language of ontology. This proved that the language used in controlled vocabularies is insufficient to identify and classify semantic information in medical narratives. Although, SNOMED CT clinical vocabulary cannot directly incorporated with medical narratives written by clinicians, still served as a useful resource to recognise the gap between controlled vocabularies and medical narratives. On the other hand, it is difficult to achieve general applicability using machine learning approach because it can only perform better in case of similar data (training and test). The overall recall, precision and f-measure for three systems are shown in Fig. 2.

5 Conclusions and Future Work

This paper presented a rule-pattern-based semantic tagger (SnoMedTagger) for the identification and classification of all possible semantic information in medical narratives. SnoMedTagger will facilitate researchers to extract semantic information from medical narratives with

the categorisation of SNOMED CT standard semantic categories. The corpus-based rule-patterns and rule-patterns analysed by refining SNOMED CT ensure that the coverage of SnoMedTagger is not only limited to medical narratives but the framework may also be helpful for researchers to analyse the limitation of controlled vocabularies (UMLS, SNOMED CT, ICD-10, etc.) on real world datasets.

We presented the results of our system on unseen test data to prove the general applicability of rule-based SnoMedTagger and also compared the output of two systems (SVM-based system, bioportal web annotator) on the same test dataset. Reasonable accuracy was achieved on unseen test dataset but we still believe in further evaluation of SnoMedTagger on more than one dataset.

Moreover, to improve the accuracy of SnoMedTagger framework, future directions also include the investigation of rules on different test cases from real world datasets and then validation of extracted concepts by getting feedback from different domain experts. We expect to contribute our semantic tagger as open source tool for research purposes.

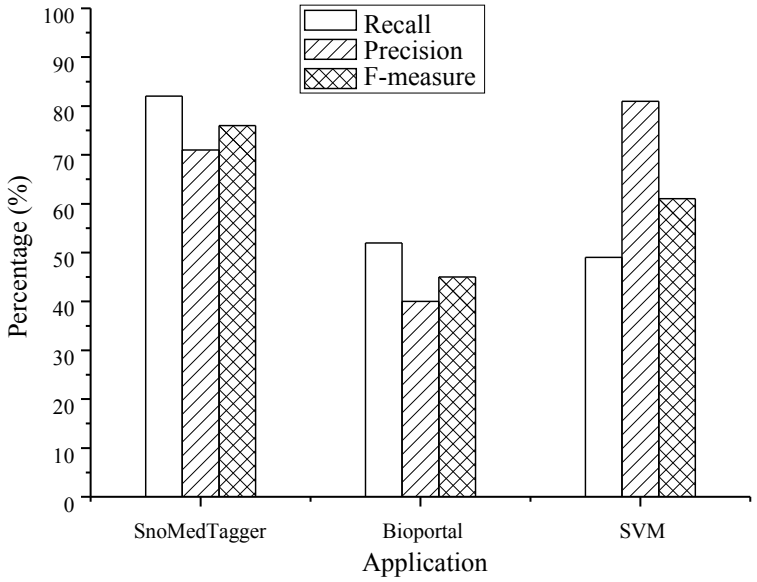


Fig. 2. Evaluation of SnoMedTagger, SVM-UM and Bioportal application against gold standard test dataset

6 References

1. Ananiadou, S., Sullivan, D., Black, W., Levow, G.-A., Gillespie, J. J., Mao, C., Pyysalo, S., Kolluru, B., Tsujii, J. & Sobral, B. 2011. Named Entity Recognition for Bacterial Type IV Secretion Systems. PLoS ONE, 6**,** e14780.
2. Ao, H. & Takagi, T. 2005. Alice: An Algorithm to Extract Abbreviations from MEDLINE. J Am Med Inform Assoc, 12**,** 576 - 586.
3. Aronson, A. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings AMIA Symposium**,** 17 - 21.
4. Cunningham, H., Mayard, D. & Tablan, V. 2000. JAPE: a JAVA Annotation Patterns Engine Second Edition ed. Sheffield: University of Sheffield.
5. Gaizauskas, R., Cunningham, H., Wilks, Y., Rodgers, P. & HumphreyS, K. GATE: an environment to support research and development in natural language engineering. Tools with Artificial Intelligence, 1996., Proceedings of Eighth IEEE International Conference 16-19 Nov. 1996 1996. 58-66.
6. Hepple, M. 2000. Independence and Commitment: Assumptions for Rapid Training and Execution of Rule-based Part-of-Speech Taggers. in Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000).
7. Hina, S., Atwell, E. & Johnson, O. 2010. Semantic Tagging of Medical Narratives with Top Level Concepts from SNOMED CT Healthcare Data Standard. International Journal of Intelligent Computing Research (IJICR), 1**,** 118-123.
8. Hina, S., Atwell, E. & Johnson, O. Enriching the corpus of Natural Language Medical narratives with healthcare data standard SNOMED CT. Corpus Linguistics, 2011 Birmingham, United Kindom.
9. Jonquet, C., Shah, N. & Musen, M. 2009. The Open Biomedical Annotator. AMIA Summit on Translational Bioinformatics. San Francisco.
10. Khare, R., An, Y., Li, J., Song, I.-Y. & Hu, X. 2012. Exploiting semantic structure for mapping user-specified form terms to SNOMED CT concepts. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. Miami, Florida, USA: ACM.
11. Li, Y. & Shawe-Taylor, J. The SVM with uneven margins and Chinese document categorization. The 17th pacific Asia Conference on Language , Information and Computation (PACLIC17), 2003 Singapore. 216–227.

12. Mehdi Embarek & Ferret., O. Learning Patterns for Building Resources about Semantic Relations in the Medical Domain. In Proceedings of LREC'2008. , 2008.
13. Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C. & Hurdle, J. F. 2008. Extracting information from textual documents in the electronic health record: a review of recent research.
14. Nadeau, D. & Turney, P. 2005. A Supervised Learning Approach to Acronym Identification. In Proceedings of Canadian Conference on AI'2005.
15. Noy, N., Shah, N., Whetzel, P., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D., Smith, B., Storey, M., Chute, C. & Musen, M. 2009a. Bioportal: Ontologies and Integrated Data Resources at the Click of a Mouse. Nucleic Acids Res.
16. Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G. & Musen, M. A. 2009b. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research, 37**,** W170-W173.
17. Ogren, P., Savova, G. & Chute, C. Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition. LREC, 2008.
18. Ohta, T., Tateisi, Y. & Kim, J.-D. 2002. The GENIA corpus: an annotated research abstract corpus in molecular biology domain. Proceedings of the second international conference on Human Language Technology Research. San Diego, California: Morgan Kaufmann Publishers Inc.
19. Pustejovsky, J., Castaño, J., Cochran, B., Kotecki, M. & Morrell, M. (eds.) 2001a. Automatic Extraction of Acronym-meaning Pairs from MEDLINE Databases.: IOS Press.
20. Pustejovsky, J., Castano, J., Cochran, B., Kotecki, M., Morrell, M. & Rumshisky, A. 2001b. Extraction and disambiguation of acronym-meaning pairs in medline. Medinfo, 10**,** 371-375.
21. Reeve, L. & Han, H. 2007. CONANN: An Online Biomedical Concept Annotator. Lecture Notes in Computer Science, 4544**,** 264.
22. Roberts A, G. R., Hepple M, Davis N, Demetriou G, Guo Y, Kola J, Roberts I, Setzer A, Trapuria A, Wheeldin B. 2007. The CLEF corpus: semantic annotation of clinical text. AMIA Annu Symp Proc**,** 625-629.
23. Schwartz, A. & Hearst, M. 2003. A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. Proceedings of the 8th Pacific Sym-posium on Biocomputing: 03-07 January 2003; Lihue, Hawaii**,** 451–462.

24. Seth, K., Bies, A., Liberman, M., Mandel, M., Mcdonald, R., Palmer, M. & Schein, A. Integrated annotation for biomedical information extraction. Proceedings of the BioLINK 2004, 2004.
25. Settles, B. 2005. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics, 21**,** 3191 - 2.
26. Stearns, M. Q., Price, C., Spackman, K. A. & Wang, A. Y. 2001. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp**,** 662–666.
27. Wang, X. 2007. Rule-Based Protein Term Identification with Help from Automatic Species Tagging. Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2007. Mexico City, Mexico: Springer-Verlag.
28. Zhenfei, J., Jian, W. & Fei, Z. Named Entity Recognition from Biomedical Text Using SVM. 5th International Conference on Bioinformatics and Biomedical Engineering, (iCBBE) 2011., 10–12 May 2011. 1-4.

SAMAN HINA
SCHOOL OF COMPUTING,
UNIVERSITY OF LEEDS,
UK
AND DEPARTMENT OF CS&IT,
NED UNIVERSITY OF ENGINEERING & TECHNOLOGY,
PAKISTAN
E-MAIL: <SCSH@LEEDS.AC.UK, SAMAN.HINA@GMAIL.COM>

ERIC ATWELL
SCHOOL OF COMPUTING,
UNIVERSITY OF LEEDS,
UK
E-MAIL: <E.S.ATWELL@LEEDS.AC.UK >

OWEN JOHNSON
SCHOOL OF COMPUTING,
UNIVERSITY OF LEEDS,
UK
E-MAIL: <O.A.JOHNSON@LEEDS.AC.UK>