Enhancing Semantic Relation Quality of UMLS Knowledge Sources

Demeke Ayele
Addis Ababa University
Addis Ababa, Ethiopia
+251-911-107468
demekeayele@gmail.com

Jean-Pierre Chevallet
University of Grenoble, France
Grenoble, France
+33687096497
jean-pierre.chevallet@imag.fr

Million Meshesha
Addis Ababa University
Addis Ababa, Ethiopia
+251-911-318062
meshe84@gmail.com

Getnet Kassie
Addis Ababa University
Addis Ababa, Ethiopia
+251-911-245861
getnetmk@gmail.com

ABSTRACT

The quality of semantic tuples (semantic triples forming *subject-predicate-object*) has significant impact in most text mining and knowledge discovery applications. The practical success and usability of these applications momentously depends on the quality of the extracted semantic triples. Most biomedical semantic resources have been developed for different contexts focusing on the structural representation but with less attention on the acceptability and naturalness of the individual semantic triples. In this article, we presented an integrated approach for enhancing the quality of semantic tuples in the UMLS knowledge sources. The approach is based on the integration of three existing auditing techniques: avoiding redundant classifications of semantic concepts, reducing hierarchical and associative relationship inconsistencies. We evaluated the approach based on the number of identified wrongly assigned concepts and inconsistent relationships obtained. The quality of each semantic triple is evaluated based on the acceptability and naturalness of the semantic tuples. The evaluation shows promising results. In the evaluation, we have extracted 10,082 semantic triples randomly from UMLS and obtained 5646 taxonomically and 4436 non-taxonomically related semantic triples. 826 concepts are found redundantly classified and 352 are found hierarchically inconsistent. In non-taxonomic semantic triples, out of 4436, 726 are found to be inconsistent. The quality (acceptability and naturalness) of each semantic triples of the first 100 are also evaluated using domain experts. The Cohen's kappa coefficient is used to measure the degree of agreement between the annotators and the result is promising (0.8).

MEDES'12, Oct 28-31, 2012, AAU, Addis Ababa, Ethiopia.
Copyright © 2012 ACM 978-1-4503-1755-9/10/10...$10.00.

Categories and Subject Descriptors

D [**Knowledge Extraction**]: structured, semi and unstructured
D.1.1 [**UMLS**]: semantic network, Metathesaurus, taxonomic and nontaxonomic relations

D.1.2 [**Semantic Structure**]: concepts, relations, semantic tuples

D.1.3 [**Inconsistency**]: redundant classification, hierarchical and associative relations

General Terms

Algorithms, Standardization, Verification

Keywords

UMLS semantics, acceptable semantics, domain semantic, semantic tuples, knowledge extraction

1. INTRODUCTION

In biomedicine, large number of semantic resources is emerging every decade and existing ones are integrating into large scales. Though the existence of these resources has a significant impact on text mining and knowledge discovery technologies and applications, the low semantic quality (acceptability and naturalness) and specifity of these resources have imposed on the practical success of knowledge extraction and text mining applications as well as research investigations in the field [1] [6] [10] [15].

Furthermore, the usability of such applications crucially depends on the accuracy and quality of the extracted semantic triples [15] [22]. Therefore, high quality and generic semantic resource is critical for the practicality and usefulness of such applications in addition to research investigations in the field [1] [10].

Most semantic resources have been developed for specific subdomain application of biomedicine focusing on the structural representations providing less attention on the semantics of each semantic triples and the resource as a whole [1] [6] [10]. That is, the resources have been focused on specific domains' structural representations with different semantic contexts, rendering them difficult to be used in large scale knowledge discovery and text mining applications [10].

In biomedical domain, various semantic resources have been developed recently [2] [16]. They range from terminologies (e.g. UMLS [17] [18]) to ontologies (e.g. BioTop [3]). Most ontological resources contain top-level semantics of the domain resulting lack of fine-grained semantics, which are significantly applicable for reasoning and intelligent systems application [3]. The most common used resources in text mining and knowledge discovery applications, as well as researches in the field are terminologies (e.g. UMLS).

These resources are matured currently and contain fine-grained semantic triples in the domain. The UMLS semantics, for example, have been used to measure the correctness and usefulness of extracted semantic triples in semantic relation extraction [1] [6] [10]. As UMLS is the integration of many terminologies in biomedicine [17] [18], it is a widely accepted semantic resource to represent the biomedical domain. Therefore, it has richer domain semantic content than other terminological resources in biomedicine yet.

However, most terminological resources are developed using experts for specialized applications [17] [18]. This makes semantic triples at different resources to have various semantic interpretations and views, which leads to many inconsistencies and ambiguities in their domain representation formalism [11-13]. This problem is intensified if the resources are combined (e.g. UMLS) as it tries to integrate the different views and interpretations of the semantic triples, which significantly affects the correctness and quality of the semantic tuples [11] [14].

Auditing systems have been developed to asses the correctness, inconsistencies and ambiguities in the semantic resources and to suggest corrective measures for enhancing the semantical and structural representation of them. The works in [20-23], for example, are developed to asses the inconsistencies and ambiguities inherent to UMLS knowledge sources. But, while auditing systems have made large contributions in detecting the inconsistencies and ambiguities, the large volume of the UMLS and many inconsistencies and ambiguities in it [11] [14], makes to have inherent problems in this respect yet.

Consequently, these lead to incorrect interpretation of the semantic triples, resulting less correctness and quality of the semantic tuples. Furthermore, the auditing systems developed yet have been focused on the identification of one type of inconsistencies or ambiguities, which results the detection of either hierarchical or associative inconsistencies or other.

However, for high quality and correctness of the semantic triples, possibly all type of inconsistencies and ambiguities should be identified and avoided, which leads to the need of an integrated approach in identifying and circumvent multiple types of inconsistencies and ambiguities in the UMLS resources.

In this context, an integrated approach is required for enhancing the quality and correctness of semantic triples in biomedical semantic resources such as UMLS. Such approach needs a comprehensive semantic analysis of the UMLS knowledge resources (e.g. semantic network and Metathesaurus) to guarantee the unambiguateness, correctness, consistency and quality of the semantic triples in them [11] [14].

Before integrating the three auditing techniques, semantic analysis is made in four perspectives where most inconsistencies are assumed to happen in the UMLS [20-23]. First, classifications of Metathesaurus concepts into semantic classes (semantic types) are assessed for ensuring the correct assignments of concepts in UMLS [12]. Second, the hierarchical semantic inconsistencies held among semantic concepts in UMLS knowledge sources are assessed to guarantee the hierarchical consistencies [8].

Thirdly, the associative semantic inconsistencies among UMLS concepts, which are related non-taxonomically, are assessed to guarantee associative relations consistency among semantic triples in the UMLS semantic network and the corresponding triples in the UMLS Metathesaurus [5] [23]. Lastly, align the UMLS knowledge sources' semantics for integration.

In this article, an integrated approach is presented to enhance the semantics of UMLS into consistent and domain expert acceptable semantic with assessment and analysis of the UMLS knowledge sources. Existing algorithms are applied to assess and identify the semantic inconsistencies in UMLS knowledge sources and align the knowledge sources in the UMLS. We have evaluated our approach based on the number of identified inconsistencies and redundancies in UMLS, and the semantic qualities and correctness of each semantic triple.

The quality and correctness of the semantic triples are evaluated using expert judgments. Each semantic triple are transformed into human readable form and presented to subject. The subject judges the semantics of the triple based on the correct semantic arrangement of subject-predicate-object triplets. The subject rates each semantic triple by providing judgmental value of 1 or 0, where 1 is acceptable and 0 is unacceptable.

The degree of agreement between two subjects is computed using Cohen's kappa coefficient (k). In this way, the quality and correctness of the semantic triples are judged by two domain subjects, the result obtained is promising. Finally, the results are discussed and concluded with future works.

2. BACKGROUND

According to literature surveys in [13], several semantic resources have been developed in biomedicine. Generally, the resources can be categorized into terminologies and ontologies based on the semantic content they have [13]. UMLS, one of the terminological resources, is an integration of more than 150 biomedical vocabulary sources into its Metathesaurus. The UMLS Metathesaurus consists of more than 3 million concepts and their relationships [17] [18].

The UMLS has three major components: the Semantic Network (SN), the Metathesaurus (MT), and the Specialist Lexicon (SL). The semantic network is a high level semantic abstraction of the Metathesaurus. It has 135 semantic classes (types) and 54 semantic relationships. Every semantic concept in Metathesaurus is classified under at least one semantic type in the semantic network. This links the semantics of Metathesaurus to that of the semantic network.

That is, the semantic triples forming *subject-predicate-object* triplets in the Metathesaurus are semantically linked to that of the semantic network in categorizing the UMLS MT concepts into semantic classes. The semantic concepts in Metathesaurus are represented with synonymous terms from multiple vocabulary sources. In this context, the UMLS knowledge sources, the semantic network and Metathesaurus, are semantically connected to structure the semantics of biomedicine.

The integration of several vocabulary sources into UMLS has been made using experts with a goal to create a semantic link among the different biomedical resources by preserving the semantics views and terms in the original resources. This leads the UMLS to have inherent inconsistency and ambiguity problems in its semantic content [11-21].

According to empirical results in auditing systems on UMLS [4] [5] [7-9] [11] [12] [14] [20] [21] [23] [25], the major sources of these problems are: 1) Due to errors made by experts in the integration process; 2) inconsistencies and ambiguities that arise in the process of preserving the different semantic views and contexts of the original sources in the integration; 3) redundant assignments of semantic types [24] and cyclic definitions [21].

Erdogan et al. [11] quantified the semantic inconsistencies in UMLS concepts from the perspective of their hierarchical relationships and showed how inconsistent concepts can help reveal erroneous synonymy relationships. The study evaluates consistency by comparing the semantic groups of hierarchically related pair of concepts.

As a result, 81,512 concepts were found to be inconsistent due to differences in semantic groups of a concept and its parents. Morrey et al. [20], presented Neighborhood Auditing Tool (NAT), which facilitated the UMLS auditing tasks. It supports neighborhood based auditing, where an auditor concentrates on a focused concept and one of a variety of neighborhoods of its closely related concepts. It also allows an auditor to display knowledge from the two UMLS knowledge sources.

Cimino [7] developed UMLS semantic based techniques to audit the UMLS Metathesaurus for identifying possible inconsistencies. The result of the study showed that out of 57,592 concepts with multiple semantic types, 3.2% were judged ambiguous. Keyword analysis showed 7121 pairs of interchangeable terms. Using the keyword pairs, 5031 pairs of potentially redundant concepts were suggested, of which 65.1% were judged to actually be redundant.

Review of the 100,586 parent–child relationships revealed that 0.54% of them are incorrect. Review of the 219,664 Other Relationships (RO) (see in table 1 below) suggested 1299 places in the Semantic Network where relations between pairs of semantic types could be added.

Auditing methods can be classified as logic and non-logic based [9] [21]. While the logic based methods have been better performing, the semantic structure of UMLS is not consistent with it [9] [21]. The non-logic based methods [4] [7] [8] [11] [12] [20] [21] [23] detect and avoid semantic inconsistencies and ambiguities based on semantical and structural properties of the UMLS semantics and fix the problems manually.

The methods detect redundant assignments, hierarchical and associative semantics inconsistencies, and hierarchically circular relationships. The purpose of the methods is to enhance the correctness and semantic quality of the UMLS knowledge sources. More comprehensive literature survey about auditing methods can be referred in [25].

Some semantic relation extraction systems, in biomedicine, have also used the UMLS semantics for evaluating the correctness of extracted semantic triples. For example, in [10] the correctness of the extracted semantic triples is evaluated against the semantics of the UMLS semantic network. Accuracy is measured in terms of the number of concepts extracted compared to those actually exist in a sentence and the quality of the tuples was compared to manually generated semantic structures.

According to the paper, a semantic triple is correct if both biomedical concepts exist in a sentence and the semantic triples of the concepts are according to manually constructed representations.

In this context, the correctness and quality of extracted semantic triples is crucially dependent on the correctness of UMLS semantic triples, which is not always true as UMLS has many errors, inconsistencies and ambiguities inherently. In addition, manual construction of a semantic structure is very limited and consumes more time and effort.

3. MATERIALS

The UMLS knowledge sources, Semantic Network (SN) and Metathesaurus (MT), are used as baseline semantic resources to enhance the consistency and quality of the Unified Medical language System (UMLS) semantic tuples under a formal representation of *object-attribute-value* triplet. According to studies in [17] [18], UMLS combines many medical vocabularies and provides a semantic mapping structure among them. It is composed of two main knowledge components, the metathesaurus and the semantic network. The semantic structure in the UMLS is inherently related to the semantic structures of its semantic knowledge sources.

The semantic network consists of 135 semantic types that have been aggregated into a set of 15 semantic groups to reduce complexity [19]. For example, the semantic type ***Finding*** and ***Pathologic Function*** belong to the semantic group ***Disorders***. The semantic types are linked using 54 semantic relationships. For example, the semantic type ***Body Part, Organ, or Organ Component*** is associated with the semantic type ***body substance*** by the semantic relationship *location_of*. The semantic type ***dysfunction*** is related to the semantic type ***biologic function*** hierarchically, *isa*.

In semantic network, semantic types are related taxonomically in a single inheritance relationship. The hierarchy is rooted at two nodes, the entity and event. Along the hierarchy, the associative relationships defined in the ancestor semantic types are easily inherited by the decedent semantic types unless otherwise the inheritance is blocked explicitly. If a relationship can not be inherited, it is blocked in two ways.

The first is inheritance blocking (B), to mean the relationship cannot be inherited by the descendant semantic types. There are also cases where semantic relationships are Defined but Not Inherited (DNI). The relationships are used only in the defining semantic types but not inherited by its decedents.

The semantic types in the semantic network and concepts in the Metathesaurus are related using categorization links. These links are assumed as hierarchical (isa) relationships. Intuitively, it is assumed that a semantic relationship defined between two semantic types is also inheritable between pair of concepts categorized in the two semantic types.

For example, the relationship ***affects*** is defined between ***Acquired Abnormality*** and ***organism function*** as (*acquired abnormality, affects, organism function*). If it is inheritable, the relationship or its decedents is inherited between concepts categorized in *Acquired Abnormality* (e.g. C0001168) and *Organism Function* (e.g. C0000934) as (C0001168, affects/causes/induces, C0000934).

Though difficult and challenging, associative relationships (e.g. *affects*) defined between two semantic types in the semantic network can be inherited by a pair of concepts in the Metathesaurus categorized under the corresponding semantic types [23]. They are not explicitly defined among concepts, which results the requirement of mapping the semantic network relationships or their decedents.

A deeper explanation of a signature based mapping strategy of the two knowledge sources is described by Vizenor et al in [23]. Furthermore, some Metathesaurus relationships can't map to the existing semantic network relationships, which also results the need of defining additional semantic network relationships.

In this study, we have considered relationships that exist in the Metathesaurus, specifically RO (relationships other than hierarchical, sibling and synonymous), and only the existing semantic network relationship mapping are made. Concepts in the Metathesaurus are groups of similar terms from the various source vocabularies. These terms create linkage to the SPECIALIST Lexicon, which in turn enables to create linkage to domain texts. Similarly, relationships between concepts can be mapped among terms and in turn between span of texts in the discourse.

The UMLS semantic group, semantic network and Metathesaurus files and the semantic binding among them are considered as a semantic knowledge source except co-occurrence relationships.

4. METHOD

An approach is proposed to enhance the quality of semantic relations forming subject-predicate-object triplets in the Unified Medical Language System (UMLS) knowledge sources. The approach is based on improving the consistency and quality of semantic tuples in the UMLS, and representing them in the form of *object-attribute-value* or *subject-predicate-object* triplets*.*

For example, the higher order taxonomic semantic structure (*fatal heart, heart, body part organ or organ component, fully formed anatomical structure, anatomical structure)* can be represented as a set of semantically consistent subject-predicate-object triples as {(*fatal heart, heart*), (*heart, body part organ or organ components*), (*body part organ or organ components, fully formed anatomical structure*), (*fully formed anatomical structure, anatomical structure*)}.

For each semantic tuple, the *object/subject* and *value/object* are semantic types, semantic concepts or semantic atoms. The *attribute/predicate* is the semantic relationship defined/inherited between semantic types, semantic concepts or between semantic types and concepts. For example, in the semantic triple (***pharmacologic substance****, treats,* ***pathologic function***), ***pharmacologic substance*** is the *object/subject* and ***pathologic function*** is *value/object* while **treats** is *attribute/predicate*.

The approach is based on two general strategies to improve the semantic consistency and quality of the UMLS knowledge sources. The first strategy is to enhance the consistency and quality of the hierarchical semantic structure of the UMLS. The strategy identifies and amends redundantly classified concepts and hierarchical semantic inconsistencies to improve the quality of taxonomically related semantic tuples. The second strategy is also to improve the non-taxonomic semantic structures of the UMLS knowledge sources.

The strategy identifies and assesses non-taxonomically inconsistent semantic triples in the knowledge sources and avoids the inconsistencies, if any. The notations C=concept, T=semantic type, G=semantic group, R=relationship, D=inheritable, B=Blocked, DNI=Defined but Not Inheritable are used, henceforth.

4.1 Taxonomic Semantic Tuples

In the Unified Medical Language System (UMLS) knowledge sources, a semantic relation starts at the semantic context of the Metathesaurus terms, which we referred as semantic atoms, hereafter. Taxonomically related semantic triples are constructed by extracting the semantic atoms, concepts, types and groups that are related hierarchically.

In constructing the taxonomic semantic triples, the semantic types in each semantic group, the semantic concepts in each semantic type, and the semantic atoms in each semantic concept are extracted and transformed into a set of subject-predicate-object triplets.

Applying the proposed technique transforms the taxonomic semantic structure of the UMLS knowledge sources (*heart valves, heart, mediastinum, body part organ or organ components, fully formed anatomical structure, anatomical structure*) into a set of semantic triples {(*heart valves, heart*), (*heart, mediastinum*), (*mediastinum, body part organ or organ components*), (*body part organ or organ components, fully formed anatomical structure*), (*fully formed anatomical structure, anatomical structure*)}.

In assessing the acceptability and consistency of taxonomically related semantic triples, we divide the UMLS hierarchy into three layers: the semantic group, the semantic type, and the semantic concept. Extracting the semantic triples in the semantic group and semantic type layer is straightforward. That is, the semantic relationship among semantic group and semantic type is flat but the relationship of the semantic type to that of semantic group is hierarchical.

The general formalism is <***semantic type***>, <***ISA***>, <***semantic group***>. For example, the *Amino Acid, peptide or Protein* is a semantic type that has a narrower semantics to that of the semantic group *chemicals and Drugs*. The semantic triple is (*Amino Acid peptide or Protein, ISA, Chemicals & Drugs).*

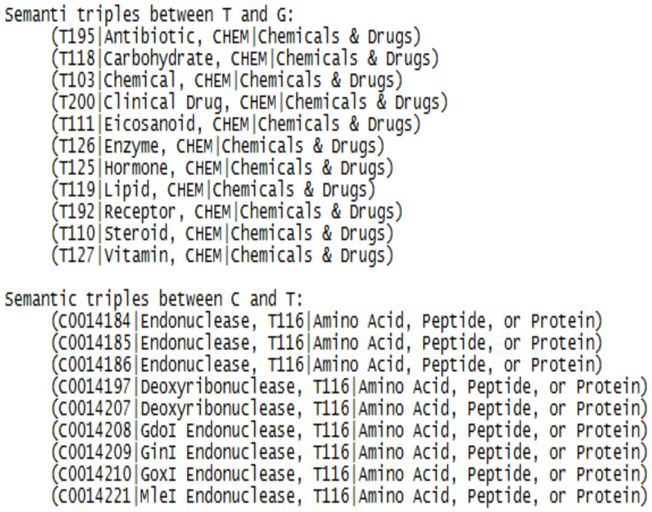
However, constructing the taxonomic semantic triples, which binds the semantic concepts and semantic types are nontrivial. In such semantic triples, the subjects are semantic concepts and the objects are semantic types forming the triples having a general formulation of (<*semantic concept*>, *<ISA>, <semantic type*>). Constructing the semantic triples in the concept layer is also nontrivial. The semantic triples have at least two inconsistency problems: redundant classification of concepts and hierarchically inconsistent semantic relationships. In section 4.3, a method is presented to avoid such problems.

As pointed out previously, the construction of taxonomically related semantic triples is straightforward. This is because the taxonomy is transitive. The semantic triples that can be derived are easily inferred from the taxonomy. Suppose a taxonomic higher order semantic relation is provided as (C_3, C_2, C_1), the taxonomic semantic triples (C_2, C_1), (C_3, C_2), and (C_3, C_1) can be inferred from the structure.

The semantic triple (C_3, C_1) is inferred from the transitive characteristics of taxonomic relationships. The fully inherited semantic network files and the hierarchical relationship of the Metathesaurus files are used to construct the taxonomic structure. The algorithm below constructs the taxonomic semantic triples.

```
Algorithm: Extracting taxonomic semantic triples
    For each sem. group, G, obtain semantic types
    For each sem. type, T, obtain semantic concepts
    Build semantic triples: ISA(C, T), ISA (T, G)
```

The snapshot shown in Figure 1 visualizes the taxonomic semantic triples resulted from the algorithm. It is constructed based on the general formulation of *ISA (T, G)* and *ISA (CT)*.



```
Semanti triples between T and G:
    (T195|Antibiotic, CHEM|Chemicals & Drugs)
    (T118|Carbohydrate, CHEM|Chemicals & Drugs)
    (T103|Chemical, CHEM|Chemicals & Drugs)
    (T200|Clinical Drug, CHEM|Chemicals & Drugs)
    (T111|Eicosanoid, CHEM|Chemicals & Drugs)
    (T126|Enzyme, CHEM|Chemicals & Drugs)
    (T125|Hormone, CHEM|Chemicals & Drugs)
    (T119|Lipid, CHEM|Chemicals & Drugs)
    (T192|Receptor, CHEM|Chemicals & Drugs)
    (T110|Steroid, CHEM|Chemicals & Drugs)
    (T127|Vitamin, CHEM|Chemicals & Drugs)

Semantic triples between C and T:
    (C0014184|Endonuclease, T116|Amino Acid, Peptide, or Protein)
    (C0014185|Endonuclease, T116|Amino Acid, Peptide, or Protein)
    (C0014186|Endonuclease, T116|Amino Acid, Peptide, or Protein)
    (C0014197|Deoxyribonuclease, T116|Amino Acid, Peptide, or Protein)
    (C0014207|Deoxyribonuclease, T116|Amino Acid, Peptide, or Protein)
    (C0014208|GdoI Endonuclease, T116|Amino Acid, Peptide, or Protein)
    (C0014209|GinI Endonuclease, T116|Amino Acid, Peptide, or Protein)
    (C0014210|GoxI Endonuclease, T116|Amino Acid, Peptide, or Protein)
    (C0014221|MleI Endonuclease, T116|Amino Acid, Peptide, or Protein)
```

Figure 1. Snapshot of taxonomic semantic triples

4.2 Non-taxonomic Semantic Tuples

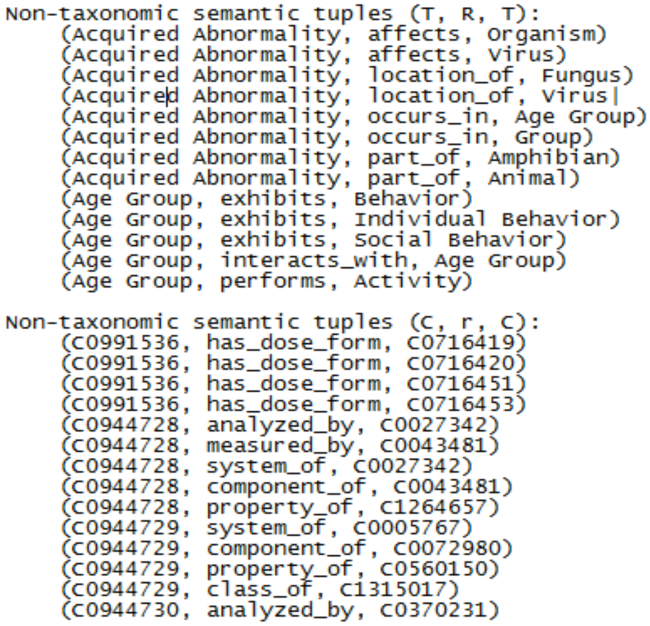
In non-taxonomic semantic triple construction, all semantic classes (semantic groups, semantic types, concepts and atoms) are considered as semantic concepts. In building the semantic triples, first the method looks for non-taxonomically related triples of a concept C_i in which C_i is the subject of the triple (*C_i,, R_{ik}, C_k*). Triples that have the same relationships (R_{ik}) and object concepts (*C_k*) are merged. Finally, only semantic triples differing with at least one of, *R_{ik}, or C_k*, , are considered useful.

Semantic relationship inheritance between semantic triples in the semantic network (*T_i, R_{ij}, T_j*) and the corresponding semantic triples in the Metathesaurus (*C_1, r_{ij}, C_j*) in which *C_i* and *C_j* are related hierarchically to *T_i* and *T_j* respectively is the mapping of *R to r*, where *r* is either the same as *R* or decedents of *R*. This mapping is valid if the inheritance of *R* is permitted (i.e. D) otherwise the mapping is invalid (B or DNI). In this article, only other relationships (RO) (other than hierarchical, sibling and synonymous) are considered to be non-taxonomic. A consistent non-taxonomic semantic triple construction is presented in section 4.3 below.

The semantic network and Metathesaurus files are used to develop the non-taxonomic semantic triples. The algorithm below illustrates the procedure to construct the non-taxonomic semantic triples.

```
Algorithm: non-taxonomic semantic triples
    For each sem. type ($T_i$), obtain ($T_i$, $R_{ik}$, $T_k$)
    For each $C_i$ in $T_i$, map R to r, obtain ($C_i$, r, $C_k$)
    Collect triples ($T_i$, $R_{ij}$, $T_j$) and ($C_i$, $r_{ij}$, $C_j$)
    Repeat from i=1 to 135, all semantic types
```

The snapshot shown in Figure 2 visualizes the non-taxonomic semantic triples resulted from the algorithm. It is constructed based on the general formulation of (T_i, R_{ij}, T_j), (C_i, r_{ij}, C_j), (T_i, r_{ij}, C_j) or (C_i, r_{ij}, T_j), where r_{ij} is narrower or similar to R_{ij}.



```
Non-taxonomic semantic tuples (T, R, T):
    (Acquired Abnormality, affects, Organism)
    (Acquired Abnormality, affects, Virus)
    (Acquired Abnormality, location_of, Fungus)
    (Acquired Abnormality, location_of, Virus|
    (Acquired Abnormality, occurs_in, Age Group)
    (Acquired Abnormality, occurs_in, Group)
    (Acquired Abnormality, part_of, Amphibian)
    (Acquired Abnormality, part_of, Animal)
    (Age Group, exhibits, Behavior)
    (Age Group, exhibits, Individual Behavior)
    (Age Group, exhibits, Social Behavior)
    (Age Group, interacts_with, Age Group)
    (Age Group, performs, Activity)

Non-taxonomic semantic tuples (C, r, C):
    (C0991536, has_dose_form, C0716419)
    (C0991536, has_dose_form, C0716420)
    (C0991536, has_dose_form, C0716451)
    (C0991536, has_dose_form, C0716453)
    (C0944728, analyzed_by, C0027342)
    (C0944728, measured_by, C0043481)
    (C0944728, system_of, C0027342)
    (C0944728, component_of, C0043481)
    (C0944728, property_of, C1264657)
    (C0944729, system_of, C0005767)
    (C0944729, component_of, C0072980)
    (C0944729, property_of, C0560150)
    (C0944729, class_of, C1315017)
    (C0944730, analyzed_by, C0370231)
```

Figure 2. Snapshot of non-taxonomic semantic triples

4.3 Consistent Semantic Tuples

Consistency can be viewed as an accurate and acceptable representation of the semantic tuples or non-redundant classification of concepts in the Metathesaurus. On the contrary, inconsistencies are resulted from inaccurate representation of the semantic network and Metathesaurus relations, inaccurate concept categorizations, and miss-categorizations of Metathesaurus concepts. Detecting and removing the redundant classifications and the inaccurate representation of semantic tuples could reduce the semantic inconsistencies.

Redundant classification occurs in cases if T_1 is decedents of T_2 and a concept C_1 is classified under T_1 and T_2. In this situation, the assignment of C_1 to T_2 is redundant. This is because it can be inferred from the assignment of C_1 to T_1 transitively. The redundant assignment (or the semantic tuple C_1 isa T_2) is removed or made implicit to make consistent. The next algorithm is a procedure to detect and remove the redundant classifications. Generally, our technique is based on the method presented by Yi et al [24].

```
***Algorithm:*** *removing redundant classifications*
    For each concept $C_i$ in MT, obtain its sem. types
    Obtain hierarchically related semantic types
    Remove the ancestor ST assignments, if any
```

Hierarchical relationship inconsistencies occur in cases where T_1 becomes an ancestor of T_2 in relationship conditions if C_1 and C_2 are related taxonomically in MT (*C_1 isa C_2*), and C_1 is under T_1 (*C_1 isa T_1*), C_2 is under T_2 (*C_2 isa T_2*). That is, T_1 must be decedent or the same as T_2 to make consistent. More details can be obtained in [8]. The next algorithm is developed based on Cimino et al [8] to detect and remove such inconsistencies.

```
***Algorithm:*** *hierarchical inconsistencies*
    For each hier related concepts, $C_i$ and $C_j$
    Obtain the semantic types for $C_i$ and $C_j$
    Remove the intersection STs of $C_i$ and $C_j$
    Verify the STs of $C_i$ are decedents of that of $C_j$
    Remove the inconsistencies, if any
```

Unlike the semantic network relationships, Associative relationships in Metathesaurus are not explicitly defined [23]. This creates difficulties in mapping the SN semantics to the corresponding MT semantics, resulting non-taxonomic relationship inconsistencies. This occurs when the semantic relationships between two semantic types, T_1 and T_2, have no direct mapping to the semantic relationships made by two semantic concepts, C_1 and C_2, which are categorized in T_1 and T_2 respectively.

For example, the semantic type ***body part, organ and organ component*** is hierarchically related to ***fully-formed anatomical structure***. The semantic type ***disease and syndrome*** is also related to ***pathologic function*** hierarchically. A semantic relationship *location_of* exists between semantic type ***body part, organ and organ component,*** and ***disease and syndrome.*** *Adrenal cortex* and *adrenal cortical hypofunction* are two Metathesaurus concepts categorized in ***body part, organ and organ component,*** and ***disease and syndrome*** respectively**.**

However, the relationship between the two concepts are not explicitly defined or inherited. In order to make consistent semantic mapping, the relationship between the two concepts should be either *location_of* or its decedents, if any.

We assumed that the inheritable relationship (R) between semantic types T_1 and T_2 or its decedents in SN are also inheritable to all concepts categorized in T_1 and T_2. This leads to develop simple algorithm to map the semantic triples in SN to semantic triples in MT. In this article, the semantic mapping considers only semantic relationships in MT. Specifically, other relationships (RO) (non-taxonomic relationships) illustrated in table 1 are considered for associative semantics mapping. After mapping the semantic relations between the two knowledge sources, manual assessment is made to assure the consistency of the mapping. In this technique, we applied the general approaches indicated in [23].

```
***Algorithm:*** *associative inconsistencies*
  For a pair of STs, $T_1$ & $T_2$, in SN related by R
      Check the inheritability (D) of ($T_1$, R, $T_2$)
      Obtain the concepts under $T_1$ and $T_2$
      Derive the r/ship, R, among concepts in $T_1$ & $T_2$
      Repeat 3$^{rd}$ step for all concepts under $T_1$ and $T_2$
```

Table 1. Metathesaurus relationships

Abbrev.	**Meaning**	**example**
CHD	Has child relationship	C_1 parent of C_2, inverse_ISA
PAR	Has parent relationship	C_1 child of C_2, ISA
RB	Has a broader relationship	C_1 parent of C_2, inverse_ISA
RN	Has a narrower relationship	C_1 child of C_2, ISA
RL	The relationship is similar or alike	C_1 alike C_2, mapping
RO	***Relationships other than CHD, PAR, RB, RN and SY***	***Associative r/ship of C1 & C2***
RU	Related, unspecified	Can be inherited from SN, T1 & T2
SIB	Has sibling relationship	C_1 SIB C_2, sistership

5. RESULTS AND DISCUSSION

The approach is evaluated by extracting and analyzing a total of 10,082 semantic triples randomly from UMLS 2010AB knowledge sources. There is no special consideration for the semantics of either hierarchical or associative relationships. The 5,646 semantic triples are found to be hierarchically related, which accounts about 56% of the total. The 4,436 semantic triples, which accounts 44% of the total, are found to be non-hierarchically (associatively) related.

This seems that hierarchically related semantic triples are provided more emphasis than associative relations. However, according to the empirical analysis, most of the semantic relationships in the Metathesaurus are hierarchical as they brought from thesauri relationships of the source vocabularies.

In an empirical analysis of the different causes of inconsistencies such as redundant classification, hierarchical and associative relationships, we have compared to the count of semantic tuples in the two semantic classes, taxonomic and non-taxonomic. This enables to forecast the trend of the possible inconsistencies in the millions of semantic triples that can be generated in the Unified Medical language System (UMLS).

In the 5646 hierarchically related concepts in MT, we have found 826 redundantly categorized concepts, which they are removed accurately. Similarly, 352 semantic concept pairs are found to have hierarchically inconsistent when compared to the corresponding semantic type pairs related hierarchically. This accounts 0.06% to the total of hierarchically related semantic triples.

In the case of non-hierarchically related semantic triples, which account 4436 tuples, it is found 726 semantic inconsistencies in mapping the semantic network tuples to that of the corresponding Metathesaurus concept tuples. Some of these inconsistencies come from lexical variations of the relationship phrases and the blocking of inheritances.

Finally, one hundred randomly selected semantic triples are presented to expert evaluators. Each semantic triple is evaluated by two evaluators and classified in either 1 (acceptable) or 0 (unacceptable). In the first evaluator, 87 are accepted and 13 are unaccepted. In the second evaluator, 93 are accepted and 7 are unaccepted. Five semantic triples are unaccepted by the first evaluator but accepted by the second. Three semantic triples are unaccepted by the second subject but accepted by the first. Twelve semantic triples are unaccepted and eighty semantic triples are accepted in common.

Cohen's kappa coefficient (k) is computed to see the degree of agreements between the two subjects where k= (pr(a)-pr (e))/(1-pr (e)). Pr (a) is the relative observed agreement and pr (e) is the probability of random agreement. The result is 0.8, which indicates better agreement between the two.

6. CONCLUSION AND FUTURE WORK

In order to utilize the UMLS semantics as a benchmark for quality semantic relation extraction in the biomedical domain, the quality and accurateness of the semantics in the UMLS knowledge sources should be analyzed and assured by domain experts and the inherent inconsistency and ambiguity problems in the UMLS need to be alleviated.

In this context, we have proposed a method for assessing semantic inconsistency problems and transforming the UMLS knowledge sources semantics into consistent and domain expert acceptable semantic tuples forming subject-predicate-object triplets. In the method, we have developed techniques to extract semantic triples from the UMLS knowledge sources and transform them into a set of semantic tuples forming *subject-predicate-object* triples. Furthermore, to assess the inconsistencies related to redundant classification, hierarchical and associative relationships, we have developed techniques based on the existing ones.

An evaluation is conducted by extracting 10,082 semantic tuples from UMLS knowledge sources, semantic network and Metathesaurus. Though the techniques can be applied on number semantic triples, the result of the evaluation is promising. Furthermore, the quality (acceptability and naturalness) of the semantic triples are also evaluated using domain experts. The Cohen's kappa coefficient (k) is used to measure the degree of agreement between the two evaluators and the result is promising (0.8).

The method developed in this article is limited to the use of the study in knowledge extraction in biomedicine. But, to utilize the full semantic potential of the UMLS, a generic and rigorous approach, which transforms its semantics to standard semantic structure and eliminate the possible inconsistencies and ambiguities are required.

REFERENCE

1. Abacha, A. and Zweigenbaum, P. 2011. Automatic Extraction of Semantic Relations between Medical Entities: A Rule Based Approach. Journal of Biomedical Semantics: Fourth International Symposium on Semantic Mining in Biomedicine, 2(2011), 1-11.

2. Bada, M. and Hunter, L. 2007. Enrichment of OBO Ontologies. Journal of Biomedical Informatics, 40 (2007), 300–315.

3. Beisswanger, E. 2007. BioTop: An Upper Domain Ontology for the Life Sciences. IOS Press, 1-7.

4. Bodenreider, O. 2001. Circular Hierarchical Relationships in the UMLS: Etiology, Diagnosis, Treatment, Complications and Prevention. AMIA, 57-61.

5. Bodenreider, O. and Burgun, A. 2004. "Aligning Knowledge Sources in the UMLS: Methods, Quantitative Results, and Applications. IMIA: Medinfo 2004, 327-331.

6. Cameron, D. 2011. Semantic Predications for Complex Information Needs in Biomedical Literature. Proceedings of the 5th IEEE International Conference on Bioinformatics and Biomedicine, 512-519.

7. Cimino, J. 1998. Auditing the Unified Medical Language System with Semantic Methods. Journal of the American Medical Informatics Association, 5 (1998), 41-51.

8. Cimino, J. and et. al. 2003. Consistency across the hierarchies of the UMLS semantic network and Metathesaurus. Journal of biomedical informatics, 36 (2003), 450–461.

9. Cornet, R. 2005. Two DL-based Methods for Auditing Medical Terminological Systems. AMIA 2005 Symposium Proceedings, 166-170.

10. Denecke, K. 2008. Semantic Structuring of and Information Extraction from Medical Documents Using the UMLS. Methods Inf. Med., 4 (2008), 425-434.

11. Erdogan, H. 2010. Exploiting UMLS Semantics for Checking Semantic Consistency among UMLS concepts. MEDINFO, 749-753.

12. Fan, J. and Friedman, C. 2008. Semantic reclassification of the UMLS concepts. Bioinformatics, 24 (2008), 1971-1973.

13. Freitas, F. and et. al. 2009. Survey of current terminologies and ontologies in biology and medicine. RECIIS – Elect. J. Commun. Inf. Innov. Health, 7-18.

14. Friedman, C. and et al. 2001. Evaluating the UMLS as a source of lexical knowledge for medical language processing. Proceedings of the AMIA Symposium, 189-193.

15. Harkema, H. and et al. 2004. A Large Scale Terminology Resource for Biomedical Text Processing. HLT-NAACL 2004 Workshop: Biolink 2004, Linking Biological Literature, Ontologies and Databases, 53-60.

16. Herre, H. and et. al. 2011. OBML - Ontologies In Biomedicine And Life Sciences. Journal of Biomedical Semantics: Ontologies in Biomedicine and Life Sciences. 2(2011).

17. Keith, E. and et. al. 1998. The Unified Medical Language System: Toward a Collaborative Approach For Solving Terminological Problems, JAMIA, 5(1998), 12-16.

18. Lindberg, D. and et al. 1993. The Unified Medical Language System. Methods of Information in Medicine, 281-291.

19. McCray, A. T. 2001. Aggregating UMLS Semantic Types for Reducing Conceptual Complexity. MEDINFO 2001, 216-220.

20. Morreya, C. P. 2009. The Neighborhood Auditing Tool: A Hybrid Interface for Auditing the UMLS. J Biomed. Inform. 42 (2009), 468-489.

21. Mougin, F. and Bodenreider, O. 2005. Approaches to Eliminating Cycles in the UMLS Metathesaurus: Naïve vs. Formal. AMIA Symposium Proceedings, 550-554.

22. Spasic, I. 2005. Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text. Briefings in Bioinformatics, Henry Stewart Publications., 6(2005), 239-251.

23. Vizenor, L. and et al. 2009. Auditing Associative Relations Across Two Knowledge Sources. Journal of Biomedical Informatics, 42 (2009), 426-439.

24. Yi, P. and et al. 2002. Auditing the UMLS for Redundant Classifications. Proceedings of the AMIA Symposium, 612-616.

25. Zhu, X. 2009. A Review of Auditing Methods Applied To The Content of Controlled Biomedical Terminologies. Journal of Biomedical Informatics, 42 (2009), 413-425.