Constructing Reference Semantic Predictions from Biomedical Knowledge Sources

Demeke Ayele[1]*, J. P. Chevallet*[2]*, Million Meshesha*[1]*, Getnet Kassie*[1]
(1) Addis Ababa University, Addis Ababa, Ethiopia
(2) University of Grenoble, Grenoble, France
{demekeayele, meshe8, getnetmk}@gmail.com, jean-pierre.chevallet@imag.fr

ABSTRACT

Semantic tuples are core component of text mining and knowledge extraction systems in biomedicine. The practical success of these systems significantly depends on the correctness and quality of the extracted semantic tuples. The quality and correctness of the semantic predictions can be measured against a benchmark semantic structure. In this article, we presented an approach for constructing a reference semantic tuple structure based on the existing biomedical knowledge sources in which the evaluation is based on the UMLS knowledge sources. In the evaluation, 7400 semantic triples are extracted from UMLS knowledge sources and the semantic predictions are constructed using the proposed approach. In the semantic triples, 87 concepts are found redundantly classified and 207 pair of semantic triples showed hierarchically inconsistent. 128 are found to be non-taxonomically inconsistent. The quality of the semantic triple is also judged using expert evaluators. The Cohen's kappa coefficient is used to measure the degree of agreement between two evaluators and the result is promising (0.9).

Construire des prévisions de référence sémantique à partir de sources de connaissances biomédicales

Les "tuples sémantiques" forment un élément essentiel à la fouille de texte et aux systèmes d'extraction de connaissances dans le domaine biomédical. Le succès en pratique des systèmes exploitant ces informations sémantique, dépend fortement de l'exactitude et de la qualité des tuples sémantiques. La qualité et l'exactitude des informations sémantiques produites automatiquement peuvent être mesurées par rapport à une structure de référence. Dans cet article, nous présentons une approche pour construire une structure sémantique à base de tuples, basée sur des sources existantes dans le domaine biomédicale. L'approche est évaluée en comparaison du méta-thésaurus UMLS. Dans une évaluation préliminaire, 7400 tuples sémantiques ont été aléatoirement extraits de UMLS et les prédictions de relations on été construites en utilisant l'approche proposée. Dans les triplets sémantiques étudiés, 87 concepts se révèlent être classés de manière redondante et 207 paires de triplets sémantiques ont une relation hiérarchique incompatible, et finalement 128 sont jugées taxonomiquement compatibles. La qualité de la relation sémantique est également jugée en utilisant des évaluateurs, experts du domaine. Le coefficient kappa de Cohen est utilisé pour mesurer le degré d'accord entre deux évaluateurs et le résultat est d'ors et déjà prometteur (0,9).

KEYWORDS: Acceptable semantics, domain semantics, knowledge extraction, semantic triples
MOTS-CLÉS: sémantique, la sémantique de domaine acceptables, extraction de connaissances, triples sémantiques

Proceedings of COLING 2012: Technical Papers, pages 133–148,
COLING 2012, Mumbai, December 2012.

1 Introduction

Semantic predictions, semantic triples, are the basic components of text mining and knowledge representation systems. Nowadays, large scale semantic prediction extraction and representation systems are increasingly emerging to sustain text mining and knowledge management systems in biomedicine. These in turn support intelligent and quality healthcare services and management (Abacha and Zweigenbaum, 2011; Cameron, 2011; Denecke, 2008; Harkema et al., 2004).

The practicality and usability of semantic relation extraction systems critically depends on the correctness, accuracy and quality of the extracted semantic predictions. The relations are formed under a general structure of *subject-predicate-object* triples (Harkema et al., 2004; Spasic, 2005), called semantic predictions hereafter. A benchmark is necessary to evaluate the accuracy and quality of the semantic relations generated by the automatic semantic relation extraction systems. This in turn improves the usefulness of the semantic predictions in the knowledge management systems (Abacha and Zweigenbaum, 2011; Denecke, 2008).

Most of the existing semantic triple extraction systems are based on either a shallow (e.g. Wordnet or Ontologies) or a narrower (e.g. terminologies) semantic resources for measuring the accuracy and quality of the extracted semantic relations (Abacha and Zweigenbaum, 2011; Cameron, 2011; Denecke, 2008). These semantic resources either lack the fine-grained semantics (e.g. Wordnet) or focus on narrower domains (e.g. terminological resources), and adopt different semantic representation contexts. This renders difficulties in the semantic resources to use them in benchmarking for independently developed semantic tuple extraction and representation systems (Denecke, 2008).

In biomedicine, various semantic resources have emerged recently (Bada and Hunter, 2007; Herre et al., 2011). They range from terminologies (e.g. UMLS (Keith et al., 1998; Lindberg et al., 1993)) to Ontologies (e.g. BioTop (Beisswanger, 2007)). Most of the ontological resources contain high level semantics of the domain (Beisswanger, 2007), resulting in lack of fine-grained semantic triples that may have significant impact on reasoning and intelligent systems application. Terminologies (e.g. UMLS) are the most common semantic resources utilized as reference in semantic triple extraction and representation because they contain the fine-grained semantic triples in a very specific domain.

For example, the Unified Medical Language System (UMLS) semantics is used to measure the correctness and usefulness of extracted semantic predictions in the work of Abacha and Zweigenbaum (2011), Cameron (2011) and Denecke (2008). According to Keith et al (1998) and Lindberg et al (1993), the UMLS is the integration of many vocabulary sources in biomedicine. It is a widely accepted semantic resource to represent the biomedicine. It has richer semantic content than other terminological resources in biomedical domain yet.

As pointed out previously, most terminological resources, however, are developed using experts for specialized application contexts in the domain (Keith et al., 1998; Lindberg et al., 1993). This makes the semantic tuples to have multiple semantic interpretation contexts and views, which leads to many inconsistencies and ambiguities in the domain representations (Erdogan, 2010; Fan and Friedman, 2008; Freitas et al, 2009).

This problem is intensified if the resources are combined (e.g. UMLS) to integrate the different views and interpretations of the semantic triples. This may significantly affects the accuracy, correctness and quality of the semantic triples (Erdogan, 2010; Morreya, 2009; Mougin and Bodenreider, 2005; Spasic, 2005; Vizenor et al, 2009).

Auditing systems have been developed to asses the semantic inconsistencies and ambiguities in biomedical semantic resources and to suggest corrective measures. For example, in (Erdogan, 2010; Morreya, 2009; Mougin and Bodenreider, 2005; Spasic, 2005; Vizenor et al, 2009), auditing systems are developed to asses the inconsistencies inherent to Unified Medical Language System (UMLS) knowledge sources.

But, while auditing systems have made large contributions in identifying the inconsistencies and ambiguities, the large volume and number of biomedical knowledge sources and many inconsistencies and ambiguities make them difficult to circumvent the inherent problems of the resources (Erdogan, 2010; Friedman et al, 2001). Consequently, using these resources as a benchmark for semantic triple extraction could lead to incorrect interpretation of the semantic triples, which results low accuracy and quality of the semantic predictions.

In this context, a reference semantic tuple structure is required to provide consistent, accurate, and high quality semantic triples for benchmarking semantic prediction extraction systems in biomedicine. The lack of a suitable gold standard reference semantic prediction structure has so far precluded the formal evaluation of semantic triple extraction systems. Most of the existing semantic extraction systems have been informally evaluated using statistical methods through error analysis. A formal evaluation requires measuring the semantic distances of extracted semantics against the benchmark semantics. That is, the spans of texts need to be mapped to concepts and their relationships in the reference semantic structure, which provides a consistent and formal representation of biomedicine.

Constructing such a reference semantic structure needs a comprehensive analysis of the biomedicine semantic knowledge resources (e.g. UMLS) to guarantee the correctness and quality of the semantic triples in them (Erdogan, 2010; Friedman et al, 2001). Furthermore, the analysis is made in perspectives where most inconsistencies and ambiguities are assumed to occur (Erdogan, 2010; Morreya, 2009; Mougin and Bodenreider, 2005; Spasic, 2005; Vizenor et al, 2009).

We have structured our semantic analysis in four perspectives before transforming the semantic knowledge sources into semantic predictions. The first semantic analysis is used to identify redundantly classified concepts to guarantee the correct assignments of concepts in the knowledge source (e.g. UMLS (Fan and Friedman, 2008)). The second semantic analysis is made for ensuring the consistency of hierarchical relationship semantics held by the biomedical knowledge sources (e.g. UMLS semantic network and Metathesaurus (Cimino et al., 2003)).

The third semantic analysis checks the consistency of non-taxonomically related semantic triples in the semantic knowledge sources (e.g. UMLS semantic network and Metathesaurus) (Bodenreider and Burgun, 2004; Vizenor et al., 2009). The fourth analysis verifies the alignment of concepts and semantic types between UMLS knowledge sources. Lastly, the UMLS semantics is transformed into a set of consistent and acceptable semantic predictions.

In this article, we presented a method to construct consistent and domain expert acceptable semantic tuple structure with assessment and analysis of the biomedical knowledge sources applied on Unified Medical Language System (UMLS). The techniques are developed to assess and identify the semantic inconsistencies and ambiguities in the biomedical knowledge sources and transform the knowledge source semantics into a set of semantic triples.

As the approach focuses at the semantic level (concept), it can be applied on languages included in the Unified Medical Language System (UMLS). That is, it can be applied for those languages in the Unified Medical Language System's knowledge sources (e.g. English and some European languages) or the language of its source vocabularies (e.g. SNOMED CT). This makes the proposed approach to have language independent nature.

The approach is based on the language model developed by the National Library of Medicine in designing the Unified Medical Language System to integrate multiple terminologies in the domain of biomedicine. It combines conceptual and lexical representations of the domain semantics. The third Unified Medical Language System (UMLS) resource component, for example, is the SPECIALIST Lexicon, which is designed to have morphological and syntactical language models.

The approach also measures the accuracy, quality and correctness of the transformed semantic tuples using experts. Each semantic tuples are transformed into human readable format and presented to experts. The experts rate the semantics of the tuples by providing judgmental value of 1 or 0, where 1 is acceptable and 0 is unacceptable. The degree of agreement between two evaluators is measured using Cohen's kappa coefficient (k). In this way, three expert evaluators judge the accuracy and quality of the semantic tuples. The result obtained is promising. Finally, the results are discussed and concluded in future works.

2 Background

According to literatures (Freitas and et al, 2009), several semantic resources have been emerging increasingly in biomedical domain. The resources may generally be categorized into lexical, terminologies and Ontologies based on the semantic content they have (Freitas and et al, 2009). For example, Wordnet could be a lexical resource, SNOMED CT or UMLS is a Terminological resource and BioTop is ontological resource.

The Unified Medical Language System (UMLS) is the largest terminological resource in the domain, which has been developed by the National Library of Medicine (NLM) since 1986 as a long term project. Currently, it is an integration of more than 150 biomedical vocabulary sources into its Metathesaurus. The Metathesaurus consists of more than 3 million concepts and their relationships (Keith et al., 1998; Lindberg et al., 1993).

The Unified Medical language System (UMLS) has three semantically correlated components that represent the biomedical domain at various level of semantic granularity. The Unified Medical Language System (UMLS) semantic network represents the high level conceptual domain representations with broader semantic classes, called semantic types. The Unified Medical Language System (UMLS) Metathesaurus also represents the fine-grained domain semantic concepts and the corresponding terms as well as relationships among concepts.

The Unified Medical Language System (UMLS) SPECIALIST Lexicon represents the linguistics knowledge sources and lexical resources. The linguistics knowledge sources include morphological and syntactic attributes of each term in the Metathesaurus. This creates a linkage to span of texts in biomedical documents.

The semantic tuples forming subject-predicate-object triples in the Metathesaurus are logically linked to the semantic network semantic tuples. In Metathesaurus, the concepts are the subjects and objects in the triple whereas the thesauri relationships are the predicates. In the semantic network, the subjects and objects are semantic classes (types) where as the predicate is the semantic network relationships.

The semantic concepts in the Metathesaurus are categorized in at least one semantic type in the semantic network. These concepts are in turn represented by several synonymous terms from multiple vocabulary sources. In this respect, the two knowledge sources of the UMLS, semantic network and Metathesaurus, are semantically linked to structure the semantics of biomedicine.

However, 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 and terms in the original resources. This leads the UMLS to have inherent inconsistency and ambiguity problems in its semantic content (Erdogan, 2010; Fan and Friedman, 2008; Freitas et al, 2009; Friedman et al., 2001; Harkema et al., 2004). According to empirical results in auditing the UMLS (Bodenreider, 2001, 2004; Cimino, 1998; Erdogan, 2010; Fan and Friedman, 2008; Friedman and et al., 2001; Morreya, 2009; Mougin and Bodenreider, 2005; Spasic, 2005; Vizenor et al, 2009), 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 views and semantic contexts of the original sources in the integration.

Erdogan et al. (2011) quantified the semantic inconsistencies in UMLS concepts from the perspective of their hierarchical relations and showed how inconsistent concepts can help reveal erroneous synonymy relations. 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. (2009), presented Neighborhood Auditing software 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 (1998) developed semantic techniques to audit 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 0.54% that was incorrect. Review of the 219,664 other relationships (RO) (e.g. see in TABLE 1 below) suggested 1299 places in the Semantic Network (SN) where relations between pairs of semantic types could be added.

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 relationship of C1 & C2
RU	Related, unspecified	Inherited from SN, T1 & T2
SIB	Has sibling relationship	C_1 SIB C_2, sistership

TABLE 1 - META relationships and their mapping

Auditing methods can be classified as logic and non-logic based (Cornet, 2005; Mougin and Bodenreider, 2005). While the logic based methods have been better performing, the semantic structure of UMLS is not consistent with it (Cornet, 2005; Mougin and Bodenreider, 2005). The non-logic based methods (Bodenreider, 2001; Cimino, 1998, 2003; Erdogan, 2010; Fan and Friedman, 2008; Morreya, 2009; Mougin and Bodenreider, 2005; Vizenor et al., 2009) 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 (Zhu, 2009).

Some semantic predictions systems, in biomedicine, have also used the UMLS semantics for accurate extraction of semantic predictions and measuring the quality of the resulting semantic propositions. For example, in 2008, Denecke the quality and correctness of the extracted semantic predictions are checked against the semantics of the UMLS semantic network in its evaluation. Accuracy is measured in terms of the number of concepts extracted compared to those actually exist in a sentence and the quality of the relation was compared to manually generated semantic structures.

In this context, a semantic structure is correct if it contains all medical concepts in a sentence and if the semantics of the concepts are according to manually constructed representations. Kilicoglu et al (2011) also developed a semantic prediction gold standard from biomedical literatures to evaluate semantic prediction systems (e.g. semRep). However, though the studies were concerned on accuracy, structural and semantical acceptability of the semantic predictions, manual construction is very limited and consumes more time and effort in large scale semantic prediction systems, which results the need of developing alternative approaches.

3 MATERIALS

The Unified Medical Language System (UMLS) Semantic Network (SN) and Metathesaurus (MT) are used as a baseline semantic resource to evaluate the approach for generating consistent and acceptable semantic predictions under a general structure of object-attribute-value triple. According to the studies in (Keith et al., 1998; Lindberg et al., 1993), UMLS combines many medical vocabularies and provides a mapping structure among them. It is composed of the semantic knowledge components, the metathesaurus and semantic network, and lexical knowledge source, the SPECIALIST Lexicon. The semantic structure in the UMLS is inherently related to the semantic structures of its semantic network and Metathesaurus. Fig. 1 below depicts the semantic relationships of the two UMLS knowledge sources.

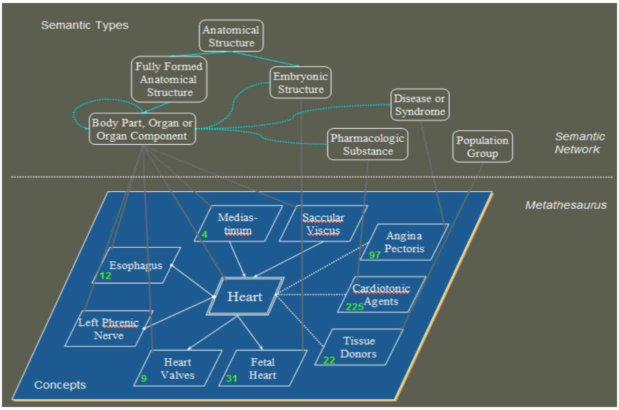


FIGURE 1- Semantic structure of the SN and MT

The upper one is the UMLS semantic network semantic types where as the lower one is the concepts in the UMLS Metathesaurus. The link between the two is the hierarchical relationship. For example, the semantic type **body part, organ and organ component** is a **fully formed anatomical structure**, which is in the semantic network. And **heart** is **body part, organ and organ component,** which is a semantic binding between Metathesaurus and semantic network.

The semantic network consists of 135 semantic types that have been aggregated into a set of 15 semantic groups to reduce complexity (McCray, 2001). 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 and concepts 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).

Fig. 2 below shows the general semantic inheritance structure between the UMLS semantic network and Metathesaurus. In the figure, the semantic types **fully formed anatomical structure** and **biologic function** is related by *location_of*. This semantic relationship can also be inherited by the descendent semantic types of **fully formed anatomical structure** and **biologic function,** which are **body part, organ and organ components**, and **diseases and symptom.** The same semantic relationship can also be inherited by the corresponding semantic concepts in the Metathesaurus between as shown in Fig. 2 below **adrenal cortex** and **adrenal cortical hypofunction**.

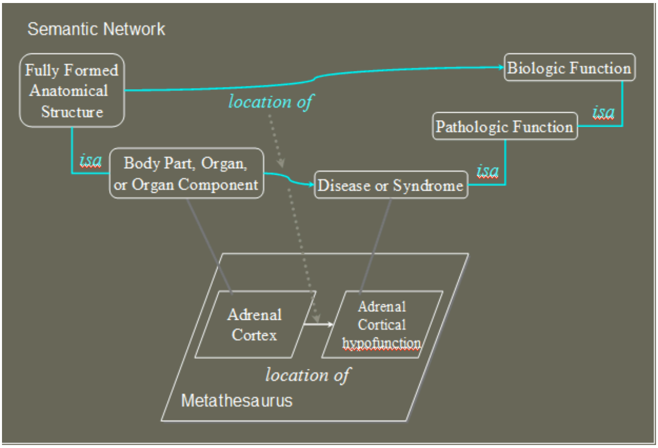


FIGURE 2 – Semantic inheritance between SN and MT

Though difficult and challenging [23], associative relationships (e.g. affects) can be inherited by pair of concepts in MT. They are not explicitly defined among concepts, which results the requirement of mapping the SN relationships. Furthermore, some MT relationships can't map to the existing SN relationships, which also results the need of defining additional SN relationships. In this study, the MT relationships considered are listed in table 1, and only the existing SN relationship mapping is made. Concepts in 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.

In this article, therefore, the UMLS semantic network, Metathesaurus and their semantic binding are used as a semantic resource except co-occurrence relationship. Within these, the SN (e.g. SRSTRE2.TXT) and MT (e.g. MRSTY.RRF, MRREL.RRF) and UMLS relationship files in addition to the semantic groups are used in constructing the semantic predictions.

4 METHOD

In this article, we proposed an approach for constructing consistent and acceptable semantic triples under a framework of *object-attribute-value*/*subject-predicate-object* triple from biomedical knowledge sources. In each semantic triple, the object/subject and value/object are either semantic types or semantic concepts or atoms. The attribute/predicate is the semantic relationship defined/inherited between semantic types or semantic concepts. For example, in the triple (***pharmacologic substance****, treats,* ***pathologic function***), ***pharmacologic substance*** is the *object/subject* and ***pathologic function*** is *value/object* while **t*reats*** is *attribute/predicate*.

In the approach, two general steps are made to complete the construction. First, all possible semantic triples in the knowledge sources (e.g. UMLS semantic network and Metathesaurus) are extracted. Second, the consistency and acceptability of the triples are assessed. The notations C=concept, T=semantic type, G=semantic group, R=relationship, D=inheritable, B=Blocked, DNI=Defined but Not Inheritable are used henceforth. Fig. 3 shown below depicts the general semantic prediction process.

4.1 Semantic Triple Extraction

In Metathesaurus, semantic relationships are based at each semantic context of the terms, which we referred as semantic atoms, hereafter. Semantic triples can be constructed at the level of semantic atoms, concepts, types and groups. That is, semantic types in each semantic group, semantic concepts in each semantic type, and semantic atoms in each semantic concept are extracted to have explicit representation of the structure. This enables to identify concepts that a semantic atom belongs, semantic types that a concept belongs, and a semantic group that a semantic type belongs. The extraction is splitted into two steps. The first is the extraction of taxonomically related semantic triples and next, the extraction of non-taxonomically related semantic triples.

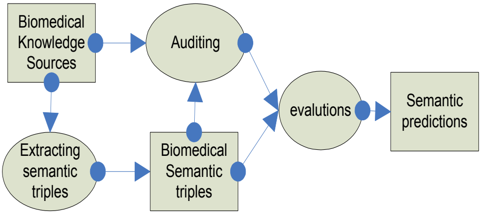

Biomedical Knowledge Sources
Auditing
evalutions
Semantic predictions
Extracting semantic triples
Biomedical Semantic triples

FIGURE 3– Semantic Prediction

Taxonomic (hierarchical) semantic triple extraction is straightforward. Because the taxonomy is transitive, relations that can be derived are easily inferred from the taxonomy. For instance, given a taxonomic hierarchy (C_3, C_2, C_1), the triples (C_2, C_1), (C_3, C_2), and (C_3, C_1) can be derived. The triple (C_3, C_1) is inferred from the transitive characteristics of taxonomic relationships. Fully inherited SN files (SRSTRE2.txt) and the hierarchical relation MT files (MRHIER.RRF and MRSTY.RRF) are used to construct the taxonomic structure. Fig. 4 below shows the result of the taxonomic semantic triple construction.

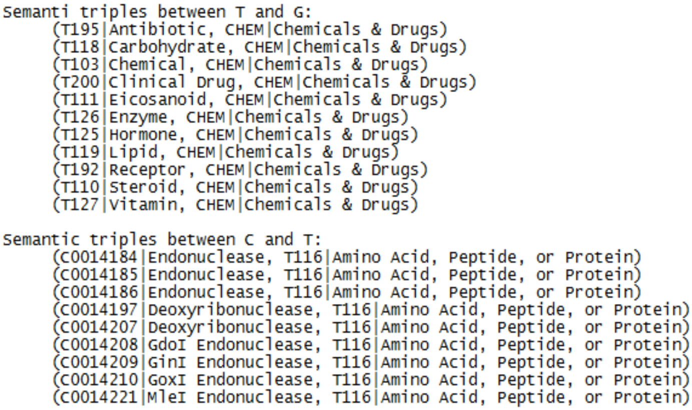
```
Algorithm: building taxonomic semantic triples

    For each sem. group, G, obtain semantic types

    For each sem. type, T, obtain semantic concepts

    For each sem. concept, C, obtain sem. atoms, A

    Build the taxonomic structure, ISA (A, C, T, G)
```



```
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 4- Snapshot of taxonomic semantic triples

In non-taxonomic semantic triple extraction, all semantic classes (semantic types, concepts and atoms) are considered as concepts. Non-taxonomic semantic triples of a concept C_i in which C_i is the subject of the triple (C_i, R, C_k) are extracted. Then, triples that have the same relationships (in R) and object concepts (in C_k) are merged. Finally, only semantic triples differing with at least one of C_i, *R, C_k,* are considered as useful.

Relationship inheritance between semantic triple in SN (T_1*, R,* T_2) and the corresponding semantic triples in MT (C_1*, r,* C_2) in which C_1 and C_2 are related hierarchically to T_1 and T_2 respectively is the mapping of *R to r*, where *r* is either 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 blocked (B or DNI). The fully inherited SN files (SRSTRE2.txt) and a MT file (MRSTY.RRF and MRREL.RRF) are used to develop the nontaxonomic structure. The algorithm below constructs the nontaxonomic propositions.

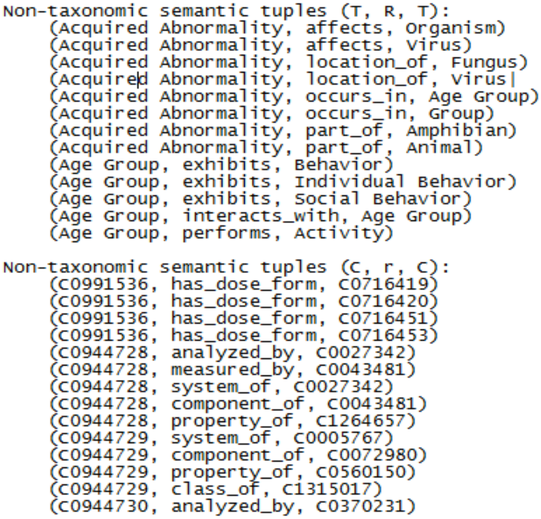
```
Algorithm: non-taxonomic semantic triples

    For each sem. type ($T_i$), obtain ($T_i$, R, $T_k$)

    For each $C_i$ in $T_i$, map R to r, obtain ($C_i$, r, $C_k$)

    Collect tuples ($T_i$, $R_{ij}$, $T_j$) and ($C_i$, $r_{ij}$, $C_j$)

    Repeat from i=1 to 135, all semantic types
```



```
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 5 - Snapshot of Non-Taxonomic Semantic Triples

4.2 Consistent Semantic Triples

Consistency is defined as accurate representation of the semantic tuples or non-redundant classification of concepts in MT. Inconsistencies are resulted from inaccurate representation of semantic network and Metathesaurus relations, and inaccurate concept categorizations. Detecting and removing the redundant classifications and the inaccurate representation of semantic tuples could eliminate 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 algorithm below is developed to detect and remove the redundancy.

```
Algorithm: removing redundant classifications
    For each concept $C_i$ in MT, obtain its sem. types
    Obtain taxonomically related semantic types
    Remove the ancestor STs, 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 in T_1 (*C_1 isa T_1*), C_2 is in T_2 (*C_2 isa T_2*). That is, T_1 must be decedent or the same as T_2 to make consistent. The next algorithm is developed to detect and remove such inconsistencies.

```
Algorithm: hierarchical inconsistencies
    For each related concepts, $C_i$ and $C_j$
    Obtain the semantic types for each, $C_i$ 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 MT are not explicitly defined (Vizenor et al, 2009). This creates difficulties in mapping the SN semantics to the corresponding MT semantics, resulting associative 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 tuples in SN to semantic tuples in MT. In this article, the semantic mapping considers only semantic relationships in MT indicated in table 1. Specifically, for example, other relationships (RO) and unspecified relationships (RU) 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.

5 Results and Discussion

The approach is evaluated by extracting a total of 7400 semantic triples from the Unified Medical Language System (UMLS 2010AB). I.e. there is no special consideration for the semantics of either hierarchical or associative relationship triples. Out of 7400, 4040 are found to be hierarchically related semantic triples, which account 55% of the total. 3360 semantic triples, which accounts about 45% 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 MT 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 each of them from the total semantic triples and to the count of semantic triples in the two semantic classes, taxonomic and non-taxonomic. This enables to forecast the trend of the possible inconsistencies in about 15 million semantic triples in the UMLS.

Out of 4040 hierarchically related semantic triples, we have obtained 87 redundantly categorized concepts, which they are removed accurately. Similarly, in the taxonomically related semantic triples, only 207 semantic triples are found to have hierarchically inconsistent in the assignments of concepts to semantic types. This account 5% to the taxonomically related semantic triples and 0.03% to the total semantic triples extracted.

In the case of non-taxonomically related semantic triples, we obtained 128 semantic inconsistencies in mapping the semantic network triples to the corresponding Metathesaurus semantic triples. This accounts 0.04% of non-taxonomically related semantic triples. 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 judged by the two evaluators and classified in either 1 (acceptable) or 0 (unacceptable). In evaluator A, 87 are accepted and 13 are unaccepted. In evaluator B, 93 are accepted and 7 are unaccepted. Five semantic triples are unaccepted by evaluator A but accepted by B. Three semantic triples are unaccepted by evaluator B but accepted by the A. 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 two evaluators 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 0.9, which indicates better agreement between the two evaluators.

Conclusion and Future Work

In order to utilize the biomedical knowledge sources as a benchmark for quality semantic prediction extraction, the quality and correctness of the semantic triples should be assured by domain experts and the inherent inconsistency and ambiguity problems need to be alleviated.

In this article, we have developed an approach for assessing inconsistency problems and transforming the knowledge source semantics to consistent and domain expert acceptable semantic triples. In the approach, we have developed techniques to extract semantic triples in the Unified Medical Language System (UMLS) and transform the triple in the form of *subject-predicate-object* triplets. Furthermore, to assess the inconsistencies related to redundant classification, hierarchical and associative relationships, algorithms are developed.

A preliminary evaluation is conducted by extracting 7400 semantic triples from Unified Medical Language System (UMLS) knowledge sources. Though the number of semantic triples considered is small, the result of the evaluation is promising. However, for accurate result and for our purpose, we will increase the number of semantic triples to one hundred thousand. Furthermore, the quality (acceptability and naturalness) of the semantic triples are also judged using domain experts. The Cohen's kappa coefficient (k) is used to measure the degree of agreement between the evaluators and the result is promising (0.9).

The approach 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 biomedical knowledge sources, a generic and rigorous approach, which transforms its semantics to standard semantic structure and eliminate the possible inconsistencies and ambiguities are required.

Reference

Abacha, A and Zweigenbaum, Z (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.

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

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

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

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

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.

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

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.

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

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

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

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

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.

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.

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.

Herre, H. and et al (2004). OBML - Ontologies in Biomedicine and Life Sciences. Journal of Biomedical Semantics: Ontologies in Biomedicine and Life Sciences. 2(2011).

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

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

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

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

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

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

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

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

Kilicoglu, H and et al (2011). Constructing a Semantic Predication Gold Standard from the Biomedical Literature. BMC Bioinformatics, 12(2011), 1-17.