Creating a Persian ontology through thesaurus reengineering for organizing the Digital Library of the National Library of Iran

Dr. Fariborz Khosravi[1]

Alireza Vazifedoost[2]

Abstract

One of the most important features in a digital library is the information retrieval system. Giving semantic search capability to such system is the goal of research in current decade. Semantic search requires establishing knowledge enabled by intelligent inferences based on an ontology. There is currently no ontology for the Persian language. Since the creation of an ontology from scratch is time consuming it was decided to reengineer a current thesaurus in the Persian language named ASFA to create the ontology using an automatic ontology learning methods. This methodology is proposed in this paper. This ontology will be used to organize contents in digital library of the National Library of Iran.

Keywords: thesaurus, thesaurus reengineering, ontology, ontology learning, information retrieval

1. Introduction

Traditional information retrieval (IR) systems which are extensively used in digital library systems, suffer from some shortcomings. These shortcomings are mainly due to two basic reasons. First, an information retrieval system cannot understand the exact intention of users and second, the system cannot understand the content of documents. In fact each keyword in a document and also in a user query may have several meanings. Therefore, just measuring the surface similarity of a document and the user's query cannot produce the best results.

To mitigate these shortcomings, some extensions have been added to

1. Assistant Prof. and Deputing Director of Research and IT of NLAI khosravi@nlai.ir

2. MSc of computer software engineering, IT Dept. of NLAI a.vazifedoost@ece.ut.ac.ir

traditional IR systems. Some of these extensions are based on using semantic properties of terms like relationships between them. A very common approach to find related terms is by using thesauri. A Thesaurus represents the knowledge of a domain with a collection of terms and a limited set of relations between them. Thesauri have been extensively used in IR system in last 2 decades.

However, there are shortcomings even with thesaurus. The main problem with thesaurus is the limitation over relation set. This sometime hides actual relations and makes ambiguous relations between terms. These ambiguous relations cause problems when using new intelligent applications, which need inferences. Another problem is the low usability in computer-based systems due to the lack of a standard representation format for them.

The ontology as a new means of representing knowledge, has received a lot of attention in recent decade. Unlike thesaurus, ontology has a non-limited set of relations between concepts and represents the explicit knowledge of a domain. Also, there are standards in computer society for representing them and these standards make them inherently usable for machines. Ontology provides an in-depth solution for considering the semantic features of documents in retrieval systems. Using ontology, documents could be semantically tagged and these tags remove ambiguities of terms. This helps the user to be able to find more precisely his/her desired content. Also, ontology provides the infrastructure for an intelligent retrieval system. These systems like Question Answering systems require inference mechanisms, which ontology fully supports.

There are many ontologies at different levels of generality in the English language. However, no Persian ontology can be found. National Library of Iran which is a pioneer in this field is creating the first Persian ontology. It would be applied in National Digital Library of Iran for knowledge organization.

Creating the ontology from scratch would take a long time to complete and when it is completed it would probably be out of date. However, one of the solutions to reduce development time is to use previously available rich knowledge representations such as thesaurus and transforming that to ontology. In our case, the ontology will be based on a previous thesaurus, named ASFA.

In this paper we describe our methodology in reengineering the ASFA thesaurus to an ontology. This project is still in its initial stages and therefore no experimental result could be provided. However, the methodology has some added value in comparison with works already done in this field. The added value of our work is two fold. First, there is no known ontology in the Persian language, and second, we use the idea of using ontology learning methods from a corpus of text materials where such works could not be found.

This paper is organized in accordance to the following sub-sections; (a) description of current thesaurus (ASFA); (b) the methodology in detail; and (c) reviews of related works and the conclusion.

2. The specifications of ASFA

ASFA (The Persian Cultural Thesaurus) is one of the most important resources for organizing bibliographical resources in the Persian language. Initial works on constructing this thesaurus began in 1991. After 4 years the initial work on the project was completed and the first edition released and now it is in its third edition. It contains domain specific terms from several domains. Also the structure of ASFA is based on international standards in thesauri building. The National Library of Iran has been using ASFA for indexing textual and non-textual materials. So far about 400,000 articles and 192,100 non-textual materials are indexed using ASFA (**12**).

We are working on one of domains in ASFA, which is the field of Library and Information Science (LIS). This work will then expand to other domains. The LIS domain contains about 800 terms and the terms are linked with BT, NT, RT, UF and USE to indicate the relationships. ASFA as a thesaurus has intrinsic problems (**15**) and these problems are as follows.

(a) Limited Semantic Coverage

The number of relationships between terms in ASFA is limited to a set of few relations. Therefore, these relationships cannot cover most real relations between concepts. On the other hand ontology requires explicit relations between concepts. The relationship A in Table 1 shows an example of such problem. "Amoozesh-e-Elme Ketabdary va Ettela'-Resani" (Library and Information Science Training) and "Daneshkad-e-Ketabdari va Ettela'-Resani"(Library and Information Science School) are two concepts in ASFA, which are assigned RT relationship. In

fact the real relation between them is a <*provide*> relationship. An ontology needs to indicate clear relationships between terms to provide background knowledge for intelligent inferences in the new semantic retrieval methods.

(b) Lack of Consistency

Another major problem with thesaurus is the lack of consistency in applying relations. The relationships, B and C in Table 1 show how using NT relationship in two cases is inconsistent. In the first case, the relationship between "Tajhizate-Ketabkhane" (*Library's Equipment)* and "Sandali" (*Chair)* is described with NT relationship. However the exact relation between these two concepts is the <*part-of*> relation. But in the second one, NT relation between "Khadamate-e-Etelaresani" (*Library's Information Services)* and "Amanat" (*Loan)* is used to show <*SubclassOf*> relationship. This kind of ambiguity makes it hard to use thesauruses in machine specific applications.

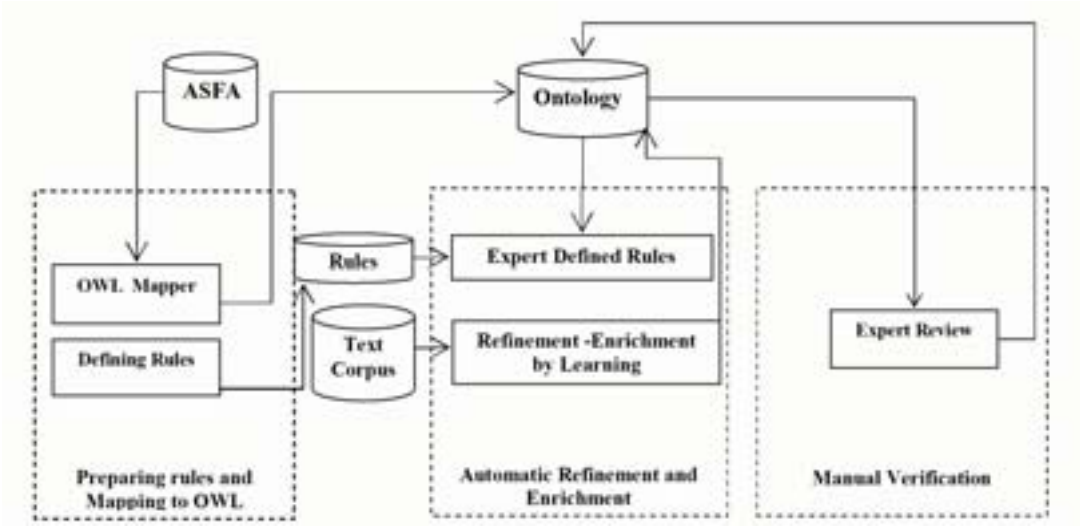
(c) Lack of a Standard Machine Readable Formats

Sharing knowledge is a goal of knowledge representation systems. Human can read two thesauri with unfamiliar formats and understand both of them but machines can just process predefined formats. For new

Table 1. Limited Semantic Coverage in Thesaurus

ASFA	Ontology
A. "Amoozesh-e-Elme Ketabdary va Ettela'-Resani" **RT** "Daneshkad-e-Ketabdari va Ettela'-Resani"	"Daneshkad-e-Ketabdari va Ettela'- Resani" **<Provide>** "Amoozesh-e-Elme Ketabdary va Ettela'-Resani"
Library & Information Science Teaching ***RT*** *Library & Information Science Faculty*	*Library & Information Science Faculty* ***<provide>*** *Library & Information Science Teaching*
B. Tajhizat-e-Ketabkhanеie **NT** Sandali	Sandali **<Part-of>** Tajhizat-e-Ketabkhanеie
Library's Equipments ***NT*** *Chair*	*chair* ***<Part-of>*** *Library's Equipmen*
C. Khadamte Ettela'-Resani-e Ketabkhane **NT** Amanat	Amanat **<Subclass-Of>** Khadamte Ettela'-Resani-e Ketabkhane
Library Information Services ***NT*** *Loan Services*	*Loan Services* ***<Subclass-Of>*** *Library Information Services*

Figure 1. The overall architecture of the system



applications in the Internet where multiple software agents have to share their knowledge, this predefined format plays a critical role. Thesauri have no such a standard format. On the other hand many works in W3C have been done to standardize ontology representations and RDF and OWL are outcomes of these works.

3. Ontology Construction Methodology

As we mentioned before, there have been several works so far on transforming thesaurus into ontology. But our approach is more similar to what is proposed by Asanee (**1**). The architecture of the system is shown in Figure 1. In this architecture, the process will be done in three steps. The first step is to prepare the structure of ontology and this is followed by defining the rules for automatic refinements. The second step is to automatically refine this mapped ontology. The goal of refinement is to clarify relationships between concepts. This job will be done through some steps which will be explained later. The last step is manual verification and enrichment of ontology by experts. Each step is further explained in subsequent sections.

(a) Mapping and Rule Definition

(i) Applying Expert Defined Rules

In the first step we transform the ASFA to an ontology structure and provide an OWL representation of it. OWL is a standard format for

representing ontology. We keep the structure of the ontology quite simple, using classes, properties and XML data types. Terms in the thesaurus would be mapped to classes and we keep the current relations between them through properties like, *NarrowerTerm* or *UsedFor* and so forth. The point is to add no interpretation to the relations at this stage. Mapping the thesaurus to the OWL format lets us use Protégé for better tracking of changes during the next steps. This is due to appropriate graphical user interface of protégé which is best designed for ontology based works. Another important point is about concepts and terms. Although a term is a lexical concept and concepts may have several lexicons, a term in the thesaurus is taken as a concept in ontology for keeping the transformation traceable.

(ii) Applying Expert Defined Rules

The second task in this step is acquiring rules. Using rules is based on the work by Asanee (**1**). We need experts to review ASFA and define rules for clarifying relations between concepts. Here we give two examples to explain the situation better. For an example of such rules consider the UF and USE relations. Although not every time but in many cases this relations in thesauri can be directly transferred to ontology as <*Similarity*> relationships. It is important that two concepts with USE and UF relations are not synonyms all the times but can be considered as similar concepts. This issue can be formulated by the following rule.

"If X and Y are two terms and X and Y **USE/UF** X then port the relation to Y <Similar> X" (1)

Apart from previous example which covers mass transformations (because it matches with a great number of source relations) other rules may be formulated that just target a portion of a source thesaurus. The next sample rule shows that a deep knowledge of current thesaurus structure and its semantic can lead us to clear relations in ontology.

"If X and Y are two terms and X Labeled as "Ket/8" (Lib/8) and Y **NT** X then port the

relation to Y <Part-of> X"

(2)

This rule explains that based on expert judgment, the concepts Y which are in NT relation with concept X with label "Ket/8" have in fact a <*Part-of*> relation with concept X.

(b) Automatic Refinement and

Enrichment

One of the main differences between a thesaurus and ontology is the degree of explicitly in their included relationships. As mentioned before, ontology requires explicit and clear relationships, where in a thesaurus do not have such a restriction. Manually clearing up ambiguous relations of thesaurus needs a lot of labor. In this context some (semi) automatic ways for clarifying these relations is needed. In the following sub-sections the steps taken by the system without human interference for clearing relations is explained.

(iii) Applying Expert Defined Rules

The previous subsystem describes about acquiring expert defined rules. An expert introduces a pattern as a rule to the system in previous section but applying it would be automatically done in the current subsystem. For an example, the application of sample rule (2) from the previous section is taken. Figure 2 shows a portion of transferred thesaurus to OWL in which sample rule (2) matches with it. In Figure 2, the relation between concept ساختمان_کتابخانه_Library_building and two other concepts, سالن_مطالعه_کتابخانه_Library_Reading_Room and تالار_مرجع_

Figure 2. The Partial Representation of ASFA in OWL Format Showing Rule Matches with a Portion of Ontology

```
<owl:Class rdf:ID="سالن_مطالعه_کتابخانه_Library_Reading_Room">
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Close_Resource</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:ID="ساختمان_کتابخانه_Library_building">
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Resource</rdfs:comment>
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >11/کد</rdfs:label>
  </owl:Class>
  <owl:Class rdf:ID="تالار_مرجع_Reference_Room">
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Open Resource</rdfs:comment>
  </owl:Class>
  <owl:ObjectProperty rdf:ID="Part-Of">
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#تالار_مرجع_Reference_Room"/>
          <owl:Class rdf:about="#سالن_مطالعه_کتابخانه_Library_Reading_Room"/>
        </owl:unionOf>
      </owl:Class>
    </rdfs:domain>
    <rdfs:range rdf:resource="#ساختمان_کتابخانه_Library_building"/>
  </owl:ObjectProperty>
```

Reference_Room is NT. This relation is ported to OWL representation with no change. However because the concept ساختمان_کتابخانه_Library_building is labeled with کت/11 (Ket/11) and the relationship between these concepts is NT we can transform the NT relationship to Part-Of as is suggested by the rule in figure 3.

The label (کت /11) Ket/11 is related to a kind of labeling used in ASFA. In this labeling system, every term in ASFA is tagged with a label showing its position in the thesaurus. Also after porting thesaurus to ontology representation a concept in ontology can be identified by means of it. Although each rule is quite clear and there is no need to essentially use this label for identifying concepts.

(iv) Refinement and enrichment by learning

Apart from using expert-defined rules which is proposed by previous works in thesaurus reengineering domain, the current suggested solution is based on ontology learning techniques. Ontology learning targets finding methods to extract knowledge from large text corpora in ontology format. The main idea is that large amounts of knowledge have been written over time by experts where this knowledge is the basic need for constructing ontology.

Figure 3. The Refined Relations are Highlighted

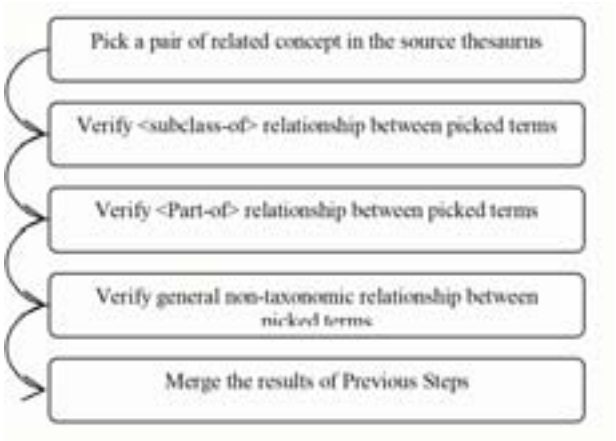
```
<owl:Class rdf:ID="سالن_مطالعه_کتابخانه_Library_Reading_Room">
   <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
   >Close_Resource</rdfs:comment>
 </owl:Class>
 <owl:Class rdf:ID="ساختمان_کتابخانه_Library_building">
   <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
   >Resource</rdfs:comment>
   <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
   >11/کت</rdfs:label>
 </owl:Class>
 <owl:Class rdf:ID="تالار_مرجع_Reference_Room"/>
 <owl:ObjectProperty rdf:ID="NarrowerTerm">
   <rdfs:domain rdf:resource="#ساختمان_کتابخانه_Library_building"/>
   <rdfs:range>
     <owl:Class>
       <owl:unionOf rdf:parseType="Collection">
         <owl:Class rdf:about="#سالن_مطالعه_کتابخانه_Library_Reading_Room"/>
         <owl:Class rdf:about="#تالار_مرجع_Reference_Room"/>
       </owl:unionOf>
     </owl:Class>
   </rdfs:range>
 </owl:ObjectProperty>
```

Therefore in analyzing the current corpora by machines it is possible to extract a portion of that knowledge for ontology creation. The idea is to use these techniques for refining the current relations ported from the thesaurus to the ontology format. The ontology is enriched with new relation by using some language dependant techniques for finding and verifying taxonomic and non-taxonomic relations. There are two notable issues when talking about ontology learning in the context of the current problem. First is the probability of using present knowledge of thesaurus for optimizing learning methods and second is the need for adapting language dependent techniques of ontology learning. In fact these techniques are mostly designed for English language and not Persian language. For ontology learning applications we need a text corpus. This corpus has to contain documents in the domain of interest which the ontology is made for. We use the collection of a journal named *Faslanme- Ye Ketab* (the quarterly journal of the National Library and Archives) which is publishing in the domain of library and information studies in Iran.

Techniques for refinement and enrichment of ASFA relations are the same. However, in refining relations, no new concepts are added and current relations are cleaned. Enrichment is done in two ways, firstly by adding new concepts and secondly by adding new relations. At this stage of the project no new concept is added to the ontology. Instead, to enrich the ontology more relations between concepts are added. In the relation refinement process each pair of related terms in ASFA is considered. The goal is to suggest a new descriptive and explicit relationship instead of current ambiguous relation. These new relationships can be divided into two groups. The first group contains taxonomic relations, which are *subclass-of* and *part-of* relations. The second group is called non-taxonomic relations and includes every imaginable relation between each two concepts such as causal, possession, similarity and so forth.

Currently, many works have been done in extracting taxonomic relations from text corpus, however non-taxonomic relations are more complex and finding them is more complicated. Three methods for extracting taxonomic relations is proposed and one for non-taxonomic relations. The number of methods used is restricted because the project is in its initial stages and the

Figure 4. The General Steps Taken in Relation Refinement Process



effects of using these methods need to be evaluated. The selected methods will be briefly explained in next subsection.

Before explaining the methods used for ontology learning, an introduction to the general process of relation refinement is provided. When refining relations it is already known which concepts exists in the process and which ones in ASFA are related together. So the first step in ontology learning is taken that is, finding the concepts specific to a desired domain. The remaining work focuses on adjusting a new label for current relations in thesaurus to make them acceptable in the ontology. Therefore, the remaining process of ontology learning involves finding explicit relations between these already known related concepts.

General steps taken for refining begin with picking a pair of related concept from the ASFA thesaurus as shown in Figure 4. This is followed by finding a relation label between this pair by applying each selected ontology learning method over the corpus. Each method will suggest a different label for a new relation between these concepts and also a confidence value between 0 and 1. For each method, the way for measuring confidence value differs from others. A preferred label among four generated label will be chosen based on expert judgment.

(v) Learning method for Subclass-Of relations

- Using NP rules

The first method for refining Subclass-Of relations is using the

structure of noun phrases. Using this structure rules are created for identifying <Subclass-of> relations. A sample of these rules which is based on noun phrase analysis and is used to analyze the surface form of a compound terms' head word is given. This rule states that if the head word of a term has the same surface form as its broader term, the system can guess there is a "subclass-of" relationship between them. Although this is not always true, but can give a good suggestion. For an example consider the following Persian language grammar for simple NP (Megerdoomian, n.d).

NP -> *head modifier*

Where the head is a noun and modifier is as shown below.

modifier -> (Adverb) Adjective

The example below represents a simple noun-phrase.

Ketab-khane (head)
(Libraries)

Then consider this relation from current ASFA thesaurus:

Ketab-khane ***BT*** *Ketab-khane-e melli*
(Libraries **BT** National Libraries)

The highlighted sections, show matching the surface form of the head words in two nounphrases in BT relation. Then we can conclude:

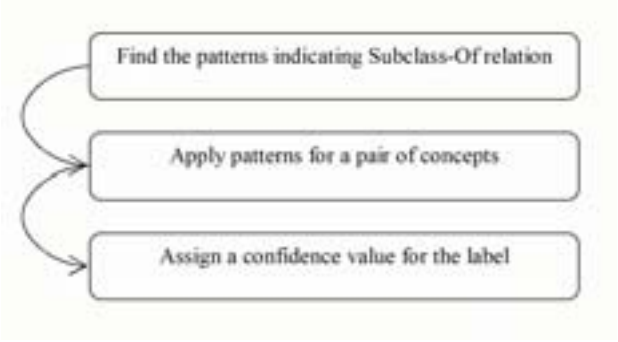
Ketab-khane-e melli <subclass-of> *Ketab-khane*

Although the output in this sample is a correct relationship but it is not always the case. Nounphrases are very ambiguous in Persian language and sometimes the above mechanism does not work correctly. More ever there are several kinds of noun-phrases in Persian which the mechanism should be adopted for each case. A study by Velardi (2006) uses Wordnet to increase the corrections of such relations. However, there is no upper level ontology like Wordnet in Persian and as such suggestions for <*subclass-of*> relations and human judge will verify those relations at a later stage (**17**).

- Using Patterns

The other method used for refining <*subclass-of*> relations is based on works done by Hearst (1992; 1998). The main idea is to use patterns which indicate <*subclass-of*> relation

Figure 5. The Process of Refining Subclass-Of Relations



between concepts. For example in English if you say "*there are many libraries such as public libraries*" , the pattern " X such as Y ", indicates a Subclass-Of relationship between *library* and *public library* (**9;10**). The main process in this method consists of three steps as is shown in Figure 5.

- Finding patterns

The first step is finding the patterns which imply <*subclass-of*> relation. These patterns are not found for Persian language. This can be done by selecting two concept which is known have such a relation between them and then finding all the occurrences of them together in the corpus. This gives sentences which contain the <*subclass-of*> pattern in each language. The next step is extracting these patterns and formulating them. Table 2 shows two sample patterns.

The column *Precision* in Table 2 shows the precision of the pattern. The precision value of the pattern is based on the portion of correctly matched instance in the training corpus to all extracted relations as indicated by the following formula.

This value will be used in assigning a confidence value to each founded relation.

- Applying Patterns

After finding the patterns, it is possible to search the corpus for the selected pair of concept and find all occurrences of these concepts which match with at least one of those patterns. The count of these co-matchings is kept as well as the count of distinct patterns which create a match.

- Assigning Confidence Value

The solution used for assigning the confidence value is based on the work by Vazifedoost (**16**). For measuring

Table 2. The Sample of Patterns and Relations Inferred Through Them

Pattern	Precision	Sample
NP0 (Mesle | Manande| Shabihe) NP1, NP2,... *NP0 (like| such as) NP1, NP2*	*0.8*	**Ketabkhane-e Mesle ketabkhane-e-Melli** *Libraries like National Library*
		Suggesting: **Ketabkhane Melli <Subclass-Of> Ketabkhane**
NP1 (Va | Ya) Har NP0 Digari *NP0 (and | or) [every] other NP1*	*0.8*	**Mikroform-ha Va Har Riznegasht Digari** *Microforms and every other Micrographers*
		Suggesting: **Mikroform <Subclass-Of> Riznegasht**

$$\text{Pattern Precision(X)} = \frac{\text{Number of correctly found Relations}}{\text{Number of all found Relations}}$$

confidence value, two parameters is introduced. The first parameter is called EvidencePrecision(EP). It stands for a sum over precisions of patterns which indicate a relation. For example if a relation is indicated by two patterns with precisions 0.8 and 0.9 then EP has a value of 1.7 . The formula could be expressed as below:

$$EP(GC,Ci) = \sum_{P \in PatternSet} presition_p$$

Where *precision p* is the precision of the pattern which indicates a relation.

The second parameter is Evidence Count (EC). EC is similar to freq (GC,Ci) in previous approach but with a difference. EC involves the precision of the patterns as indicated by the formula below:

$$EC(GC,Ci) = \sum_{p \in PattenrSet} precision_p \times Count_p(GC,Ci)$$

Where *precisionp* is as before and *Countp (GC,Ci)* is the count of matches between GC and Ci through pattern p. For example, consider that the relation **Mikroform <Subclass-Of> Riznegasht** is indicated 2 times by a pattern with a precision value of 1 and 1 times with a pattern with the precision value of 0.3. Therefore EC(Mikroform , Riznegasht) is equal to 2.3. Then it is assumed that the confidence value of a relation has a direct relation with these twoparameters, that is,

$$Conf(GC,Ci) \propto EP(GC,Ci), EC(GC,Ci)$$

Finally, the following formula is used to express that direct relation:

$$Conf(GC,Ci) = \frac{\log(\alpha\sqrt[2]{EP} + \beta\sqrt[2]{EC})}{Max(Conf)}$$

The factors α and β are two corpus specific parameters.

(vi) Learning Method for Part-Of Relations

Using patterns for extracting relations is not restricted to "Subclass-Of" relationship. Therefore, the approach of Charniak, Girju and Cimiano in using patterns for finding attribute-of relations is applied (**5;7;8**). The steps of works in refining "Part-of" relations are essentially the same as what was used for "Subclass-Of " relations, that is, finding patterns, applying patterns and assigning confidence value.

(vii) Learning Method for General Non-Taxonomic Relations

The last technique is used for labeling general non-taxonomic relations. These relations do not give a hierarchical structure to ontology in contrast with what taxonomic relations like <*subclass-of*> do. A concept when relates to other concept by using a verb, creates a non-taxonomic relation. Usually this verb can be any verb which conveys a correct and clear relationship between concepts. We can imagine this verb as a label over the relationship of related concepts. An example of these relations is shown in figure 6 :

The case B, shows a relationship of the ASFA thesaurus and in the case A the same concepts are shown when they are adjusted for ontology. In the case of ontology, the relationship is quite clear and meaningful in spite of general relationships as shown in the case B. Therefore, the question in this section is finding the verbs which convey the correct relations between two already related concepts in the thesaurus.

The basic idea here as stated by Kavalec is to select verbs (or simple verb phrases) frequently occurring in the context of each two related concepts in the text corpus. The *concept-verb-concept* triples are then ordered by a numerical measure and the top ones are the candidates for relation labels of the given pair of concepts. Before describing the numerical measurement, a clear meaning of co-occurrency of two concepts and a verb need to be explained. Two concepts *c1* and *c2* and the verb *v* are co-occurred if c1 and c2 both occur within *n* words from

Figure 6. General Non-Taxonomic Relation

A	**Nirooye-Ensani-e-Archive** *[Archive man power]*	<Hedayat-Mishavad-Ba > *[<Leads-By>]*	**Modiriat-eArcjhive** *[Archive Management]*
B	**Nirooye-Ensani-e-Archive** *[Archive man power]*	**RT**	**Modiriat-eArcjhive** *[Archive Management]*

an occurrence of *v.* This destination is called the neighborhood of the verb. The first step is finding all co-occurred verbs with a typical pair concept and keeping its frequency for being used in the next step, which is deciding for the best label. For this mission we use the formula proposed by Kavalek (**11**):

$$\text{Label Likehood}\,(c_1 \wedge c_2/v) = \frac{P(c_1 \wedge c_2/v)}{P(c_1/v)\cdot P(c_2/v)}$$

In this formula is the conditional frequency and described as following:

$$P(c_1 \wedge c_2/v) = \frac{Number\ of\ times\ c_1, c_2\ \ co-occured\ in\ the\ neighborhood\ of\ vreb\ v}{Number\ of\ times\ the\ verb\ v\ has\ occured}$$

The preference of verb v as a relation label for concepts c_1 and c_2 Bhas reasonably a direct relation with $P(c_1 \wedge c_2/v)$. But that has an inverse relation with statements $P(c_1/v)\cdot P(c_2/v)$ which indicate the like hood of independent occurrence of c_1 with v and also c_2 with v. In fact just the cases in which both of concepts co-occur with the verb are important not their individual co-occurrence with the verb and formula correctly reflects this.

(c) Manual Verification

By applying each of previous methods, multiple relation labels suggested by the learning methods for each pair of concepts is obtained. Finally, these results were merged and the best labels are kept. The merging will be done by an expert. An important point is that the expert has to consider the history of transformations when making his decision. This history can help us refine the rules of transformation. For example, if most of USE relations in thesaurus are directly ported to Similarity relations, it can be formulated as a rule in the system for automatically processing remaining USE relations. Therefore, it is a good idea to keep such information and give them to experts.

Considering this matter, the

transformability value of relation X in the source thesaurus and relation Y in the destination ontology as the likelihood of transferring relation X of thesaurus to relation Y in the ontology is defined. For estimation of this value, consideration of a portion the source thesaurus is defined as the training set. Consequently, new relations are manually assigned instead of current ones. Finally, for each relation X in thesaurus the probability of transferring to relation Y in ontology is as follows:

$$Transferability\ (X,Y) = \frac{Count(X \rightarrow Y)}{Count(X \rightarrow All)}$$

Where X is a relationship in the source thesaurus and Y is a relationship in the inferred ontology. Then is count $Count(X \rightarrow Y)$ of relations X which are transferred to relation Y and is count $Count(X \rightarrow All)$ of transferred X relations to any relation in the training set. For example if the relation NT/BT is transferred in 90 percent of cases to relation <*subclass-of*> in ontology then *Transferability (X,Y)* is equal to 90%.

For each pair of concept in thesaurus, all discovered relations with their statistical confidence value, is provided. This information is contained in OWL presentation of the ontology. The expert can browse the ontology with protégé and choose the best of these relations. This selection can be based on the confidence value of the automatically extracted label, transformability value of source relation in the thesaurus and expert knowledge. Expert can also remove all of discovered relation and instead use a more meaningful relation if it is possible. Using protégé allows the applications of these modifications directly to OWL file of the ontology.

4. Related works and conclusion

The idea of using thesauri as the base platform for ontology construction was investigated in several works. But the most effective ones is the work done by (Asanee, 2005) and his more automated version. These works describe the shortcomings of thesaurus as a knowledge representation for semantic information retrieval and also explain the project of transforming the AGROVOC thesaurus to an ontology (**1**). The important tip in this papers is applying some (semi) automated approaches for cleaning relationships of thesaurus.

Other works pay less attention to automated migration. The work of Bedi describes a manual transformation of a soil thesaurus to OWL format (**4**). A paper by Qin and Paling explains

the case of converting a controlled vocabulary into ontology. However, they have given little information about their methodology (**14**). Another study (**6**), explains the work done over a Chinese agricultural thesaurus. This work seems to give most attention to conversion to OWL and little work was done for refining relations. This is a problem seen in a project of migration of an art thesaurus to ontology offered (**18**). Two other systematic works by Assem and Assem et al., describe a method which considers clearing relations but no automatic transformation have been used (**2;3**).

The proposed method described in this study, used as much as possible from previous approaches. However, using a corpus as a source of knowledge in refining relations is a new idea. The paper shows that by application of ontology learning methods it is possible to help expert to better refine the relations. In fact, the suggested relations by the system can even enrich the ontology with more than one relation between each pair of concepts. For evaluating this methodology the study currently applies it to ASFA thesaurus at the National Library of Iran.

References

1. Asanee, Kawtrakul, ...[et al]. "Automatic term relationship cleaning and refinement for AGROVOC". 2005. [on-line]. Available: ftp://ftp.fao.org/docrep/fao/008/af240e/af240e00.pdf

2. Assem, M. van ...[et al]. "A method for converting Thesauri to RDF/OWL". presented at *ISWC'04, Hiroshima*, Japan, 2004.

3. Assem, M. van. ...[et al]. "A method to convert thesauri to SKOS". *In* ***Proc. Third European Semantic Web Conference*** (ESWC'06), Budva, Montenegro,11th - 14th June, 2006.

4. Bedi, P.; Marwaha, S. "Designing ontologies from traditional taxonomies". ***Proceedings of International Conference on Cognitive Science***, Allahabad, India, 2004.

5. Charniak E.; Berland, M. "Finding parts in very large corpora". In ***Proceedings of the 37th Annual Meeting of the ACL,*** (University of Maryland: June 20-26 1999).

6. Chun, Chang; Wenlin, Lu. "From agricultural thesaurus to ontology". ***5th AOS Workshop***. *(*Beijing, China: 27-29 April 2004).

7. Cimiano, P.; Volker, J. "A framework for ontology learning and data-driven change discovery". In ***Proceedings of the 10th International Conference on***

Applications of Natural Language to Information Systems, *Text2Onto.* (NLDB'05). (Alicante, Spain: 15-17 June 2005).

8. Girju, R.; Badulescu, A.; Moldovan, D. "Learning semantic constraints for the automatic discovery of part-whole relations". In ***the Proceedings of the Human Language Technology Conference (HLT) of the North American Chapter of the Association for Computational Linguistics,*** (Edmonton, Canada: May 27 - June 1, 2003).

9. Hearst, Marti. "Automatic acquisition of hyponyms from large text corpora". In ***Proceedings of the14th international conference on computational linguistics***. 1992. [on-line]. Available :http://citeseer.ist.psu.edu/hearst92automatic.html

10. Ibid. "Automated discovery of wordnet relations". In Fellbaum, Wordnet, ***An electronic lexical database,*** MIT Press, 1998.

11. Kavalec, Martin; Svaték, Vojtech. "A study on automated relation labeling in ontology learning". In ***Ontology learning from text: methods, evaluation and applications***. [S.L]: IOS Press, 2005.

12. Khosravi, Fariborz; Ghadimi, Narges. ***Trilingual cultural thesaurus (ASFA)***. 3rd P Edition. Tehran: National Library of Iran, 2005.

13. Megerdoomian, Karine. N.d. "Persian Noun Phrase". [on-line] Available: http://crl.nmsu.edu/ Research/ Projects/shiraz/ling/np.html

14. Qin, Jiab; Paling. Stephen, "Converting a controlled vocabulary into an ontology: the case of GEM". ***Information Research,*** Vol. 6, No. 2 (2001). [on-line] Available: http://informationr.net/ir/6-2/paper94.html

15. Soergel, D. ... [et al]. "Reengineering thesauri for new applications: the AGROVOC example". ***Journal of Digital Information,*** Vol. 4, No. 4 (2004).[on-line] Available: jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/

16. Vazifedoost, Alireza; Oroumchian, Farhard;Rahgozar, Masoud. "Application of taxonomic relations for finding similarity relationships in ontology learning". ***15th. International Conference of Electrical Engineering***, 2007.

17. Velardi, P.; ...[et al]. ***Evaluation of OntoLearn, a methodology for Automatic Learning of Ontologies in Ontology Learning and Population***, [S.L]: IOS press, 2006.

18. Wielinga, B. J. ...[et al] . "From thesaurus to ontology". In ***Proceedings of the 1st international Conference on Knowledge Capture***. *KCAP'01.* ACM Press, 2001.

تاریخ دریافت: ۱۳۸۶/۳/۲۰