A Framework for Schema-based Thesaurus Semantic Interoperability*

E. Francesconi, S. Faro, E. Marinai, M.A. Biasiotti, and F. Bargellini

Institute of Legal Information Theory and Techniques
Italian National Research Council (ITTIG-CNR)
{francesconi,faro,marinai,biasiotti,bargellini}@ittig.cnr.it
http://www.ittig.cnr.it

Abstract This work proposes a formal characterization of the schema-based thesaurus mapping problem as well as a specific approach within such framework on a case study aimed at mapping five thesauri of interest for European Union institutions.

1 Introduction

In the last few years accessing heterogeneous data sources in a distributed environment has become a problem of increasing interest. In this scenario the availability of thesauri or ontologies is an essential pre-condition to guarantee quality in document indexing and retrieval, therefore interoperability among thesauri is important to guarantee cross-collections retrieval quality [1]. This work proposes a methodological framework for semantic mapping between thesauri as well as a specific approach within such framework on a case study aimed at mapping five thesauri (EUROVOC, ECLAS, GEMET, UNESCO Thesaurus and ETT) of interest for the European Union institutions having only schema information available.

2 A formal characterization of the schema-based thesaurus mapping problem

Thesaurus mapping for the case-study is a problem of terms alignment where only schema information is available (Schema-based mapping) [2] [3]. It can be considered a problem where to measure the conceptual/semantic similarity between a term (simple or complex) in the source thesaurus and candidate terms in a target thesaurus. We propose to characterize the schema-based Thesaurus Mapping ($T_M$) problem as a problem of Information Retrieval ($IR$). As in $IR$ the aim is to find documents, in a document collection, better matching the semantics of a query, similarly in $T_M$ the aim is to find terms, in a term collection (target thesaurus), better matching the semantics of a term in a source thesaurus.

* This work has been developed within the tender n. 10118 “EUROVOC Studies” of the Office for Official Publications of the European Communities (OPOCE).
The \( TM \) problem can be formalized as \( TM = [D, Q, F, R(q_i, d_j)] \) where:

1. \( D \) is the set of the possible logical views of a term in a target thesaurus (documents representation in \( IR \));
2. \( Q \) is the set of the possible logical views of a term in a source thesaurus (queries representation in \( IR \));
3. \( F \) is the framework of term representations;
4. \( R(q_i, d_j) \) is a ranking function, which associates a real number with \((q_i, d_j)\) where \( q_i \in Q \), \( d_j \in D \), giving an order of relevance to the terms in a target thesaurus with respect to a term of the source thesaurus.

This framework can be implemented using RDF/OWL standards to represent concepts and relationships; in particular the standards SKOS (Simple Knowledge Organisation System) can be used.

3 A Thesaurus Mapping Case Study

According to the project specifications, a mapping between EUROVOC and the other thesauri of interest is expected. The basic mapping methodologies are applied to descriptors within corresponding microthesauri in their English version as a pivot language. The steps of the system workflow is here below described.

a) **SKOS Core transformation and terms pre-processing**

Thesauri XML proprietary formats are transformed into an RDF SKOS Core representation using XSLT techniques. To reduce the computational complexity, terms are normalized so that digits and non-alphabetic characters are represented by a special character; then stemming and stopwords elimination are performed.

b) **Term logical views in source (Q) and target (D) thesauri**

Term semantics is conveyed by its morphological characteristics, by the context in which it is used as well as by the relations with other terms. Therefore we propose to represent the semantics of a thesaurus term by (i) its *Lexical Manifestation*: a string of characters normalized according to pre-processing steps (the framework \( F \) is represented by strings and standard operations on strings); (ii) by its *Lexical Context*: a term vector \( d \) of binary entries (statistics on terms to obtain weighted entries are not possible since document collections are not available) composed by the term itself, relevant terms in its definition and linked terms of a T-dimension vocabulary (\( F \) is T-dimensional vectorial space and linear algebra operations on vectors); (iii) by its *Lexical Network*: a direct graph where nodes are terms and the labeled edges are semantically characterized relations between terms (\( F \) is the algebra operations on graphs).

c) **The proposed Ranking Functions (R)**

A ranking function \( R \) is able to provide a similarity measure between terms. \( R \) for *Lexical Manifestations*: Levenshtein distance/similarity applied on pre-processed strings normalized with respect to the longest string (therefore this measure varies in the interval \([0,1]\)). \( R \) for *Lexical Contexts*: Correlation between
such vectors, quantified as the cosine of the angle between these two vectors. \( R \) for *Lexical Networks*: Graph Edit Distance, namely the minimum number of nodes and edges deletions/insertions/substitutions to transform a graph \( g_1 \) into a graph \( g_2 \). Because of computational complexity we have considered three variants of the Graph Edit Distance: the *Conceptual similarity* expressing how many concepts two graphs have in common; the *Relational similarity* indicating how similar the relations between the same concepts in both graphs are; the *Graph similarity* [4] expressing the number of nodes and edges shared by two graphs over the number of nodes and edges in a reference graph.

d) **Ranking among candidate terms and mapping implementation**
Candidate terms of the target thesaurus are ranked according to the similarity measure values (\( sim \in [0, 1] \)) and a semantics to mapping relations is assigned using proper heuristic threshold values (\( T_1, T_2 \in [0, 1] \)) to decide exactMatch (\( sim < T_1 \)), partial match (broadMatch or narrowMatch) (\( T_1 < sim < T_2 \)) or No Match (\( T_2 \leq sim \)).

4 **Interoperability assessment through a “gold standard”**
Interoperability between thesauri has been assessed on a “gold standard” data set, namely the ideal set of expected correct term mappings. The “gold standard” produced by experts includes 624 relations (346 are exactMatch). System mapping performances have been assessed with respect to the “gold standard” using the system Recall since the automatic mapping is addressed to identify matching concepts within the system predictions, to be validated by humans. Preliminary experiments showed satisfactory performances to identify relations expressing generic association between terms (untypedMatch); good performances have been obtained as regards exactMatch relations, while the distinction between narrowMatch and broadMatch revealed a high degree of uncertainty. The proposed term logical views and related ranking functions outperformed a simple string matching between terms. In particular for EUROVOC vs. \{ETT, ECLAS, GEMET\} the Lexical Manifestation logical view and the Levenshtein Similarity ranking function gave the best results (untypedMatch Recall = 66.2%, exactMatch Recall = 82.3%), while for EUROVOC vs. UNESCO Thesaurus the Lexical Network logical view and the Conceptual Similarity ranking function gave the best results (untypedMatch Recall = 73.7%, exactMatch Recall = 80.8%).

**References**