UNIVERSITY
OF TRENTO

DIPARTIMENTO DI INGEGNERIA E SCIENZA DELL'INFORMAZIONE

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.disi.unitn.it

CREATING AND ALIGNING CONTROLLED VOCABULARIES

Ahsan-ul Morshed and Margherita Sini

August 2009

Technical Report # DISI-09-051

Also: Short version is accepted as a poster at Workshop on Advanced Technologies for Digital Libraries 2009 AT4DL 2009, 8th September 2009, Trento (Italy)

Creating and Aligning Controlled Vocabularies

Ahsan-ul Morshed
morshed@dit.unitn.com
Margherita Sini
margherita.sini@fao.org

[1] Department of Information and Communication Technology
University of Trento, Italy
[2] Food and Agriculture Organization of the United Nations (FAO)
Rome, Italy

Abstract. A vocabulary stores words, synonyms, word sense definitions (i.e. glosses), relations between word senses and concepts; such a vocabulary is generally referred to as the Controlled Vocabulary if choice or selections of terms are done by domain specialists. In our case,we create and match two controlled vocabularies by using their concept facet. This methodology is based on semantic matching which is different from the orthodox view of matching.

Key words: Ontology, Vocabulary, Thesaurus, AGROVOC, CABI

1 Introduction

There is a huge amount of information scattered on the World Wide Web. As the information flow occurs at a high speed in the WWW, there is a need to organize it in the right manner so that a user can access it easily. Previously the organization of information was generally done manually, by matching the document contents to some hierarchies. Hierarchical library classification systems (such as the Dewey Decimal Classification system (DDC)) [18] or the Library of Congress classification system [17] are attempts to develop static, hierarchical classification structures into which all of human knowledge of a specific domain can be classified [23, 1].

Another technology for web information management (which has gained widespread fame recently) is the Semantic Web, where the underlying idea is that web contents should be expressed not only in natural language but also in a language that can be unambiguously understood, interpreted and used by software agents, thus permitting them to find, share and integrate information more easily. The central notation of the Semantic Web's idea is the ability to uniquely identify resources (with URIs) and languages (e.g. RDF/S, OWL) to formally represent knowledge (i.e. ontologies, which can simplistically be considered the taxonomies of classes representing objects, and of their inter-relationships) [14, 2]. These taxonomies contain domain knowledge; the domain is represented by a set of words and phrases used to describe concepts. A vocabulary is said to

be controlled if it stores domain-specific chosen words, synonyms, word sense definitions (i.e. glosses) and relations between word senses and concepts [23]. In Controlled Vocabulary (CV), we denote the word as "words are the blocks from which sentences are made", a synonym as "a word or phrase that refers to the same concept", a sense as "a meaning of a concept" and a concept as "an abstract idea inferred or derived from specific instances".

The importance of CVs can hardly be underestimated;generally,each company or research group has its own information source e.g. databases, schemas and structures. Each of these sources has their respective set of individual CVs, creating a high level of heterogeneity. On the one hand this is desirable, as it allows the involved parties to structure knowledge in a way which best fits their needs, e.g., for specific inter-office applications. On the other hand, individuals or companies also sometimes need a unified knowledge base (made up of different information sources) in order to satisfy their goals. This source of integration process requires a mapping between different CVs. Mapping between two CVs is generally a critical challenge for semantic interoperability. These CVs are used a lots as background knowledge for this data integration [7, 4]. What is more, classifications are matched using CVs are lightweight ontologies, also called Formal classification (FC). In FC, lexical labels are translated to logical labels that remove ambiguities of natural language. For interested reading, we refer to [9, 5]. In our case, we are interested in the correspondence between concepts from two CVs, e.g., concept-to-concept mapping which includes word-to-word mapping, or synonym-to-synonym and senses-to-senses mapping. This mapping cannot be accomplished solely by a lexical comparison of two concepts using element level matcher [11, 16] that is included in SMOADistance, HammingDistance, JaroMeasure, SubStringDistance, N-gram, JaroWinKlerMeasure, and LavesteinDistance; we also need to consider the existing semantics. In light of the above discussion, the objective of this work is to determine a fully-automated mapping between two CVs and this work may be useful for navigating vocabularies, information extraction and linking information.

2 Automatic Controlled Vocabulary Creation

Some research has been done on CV construction by automatic or semi-automatic methods [10, 3]. These two methods can be categorized into two approaches [6]:

- Statistical Approach
- Linguistic Approach

In the statistical approach, terms are extracted from a document by IDF (inverse document frequency). Adapted to the controlled vocabulary construction problem, the assumption is that frequently co-occurring words with a text window (sentence, paragraph or whole text) point to some semantic cohesiveness. The co-occurrence approach needs human intervention before terms can be used for controlled vocabulary creations. From a linguistic approach, terms and their relations are based on the distributional context of syntactic unit (subject and

object) and the grammatical surrounding function these unit. For instance, suppose we have two terms "Agricultural business" and "Agricultural industry". These two terms cab be semantically mapped:

- The above word terms shared the same head or tail (i.e. agricultural)
- The substituted words have the same grammatical function (Modifier,i.e. business and industry)
- The substituted words are semantically close (i.e. business and industry).

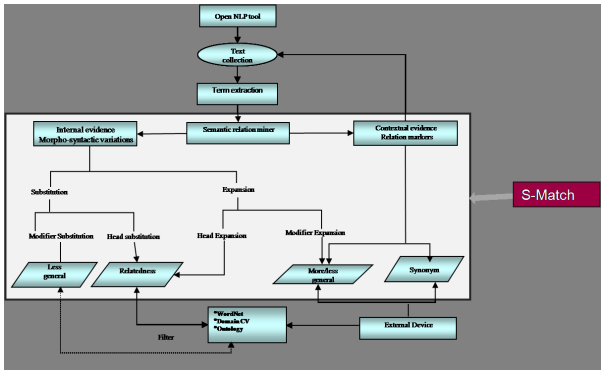

Fig. 1. Proposed Diagram of CV

The two described approaches are time-consuming and need a substantial amount of human intervention. To overcome the difficulties of these approaches, we present our proposal of controlled vocabulary in Figure 1. In our proposal we merge or combine the previously cited two approaches into one. Furthermore, we have used semantic matching algorithm to find the relations among terms, reducing time compared to the linguistic techniques. Our approach is different from other because all other existing techniques use syntactic matching techniques to find out the relatedness among the terms. Instead we use semantic matching techniques and background knowledge. Because it is difficult to find the universal background knowledge, we used WordNet [15] in order to conduct testing.

Our algorithm is defined into micro steps as follows:

Step 1: Extracting terms from a document using NLP tools.

Step 2: Building Semantic Relationships among terms and using S-match tools for calculating relatedness among the terms.

Step 3: Filtering Terms Relationships with WordNet/External Resources.

Step 4: Giving linkage information for words according to semantic similarities.

In Step 1 we take a set of documents and extract keywords using the Kea tool [12]. In Step 2 we use the Element Level Matcher from S-Match tool to

calculate the relatedness between two terms. In Step 3 we use WordNet to filter the information. After filtering, we assign to each keyword according to semantic semilaries. This work on automatic CV creation presented above is still on going: we have presented the general idea with a diagram (Figure 1) and description of the algorithm, but more work would need to be carried out in order to extend the testing.

In the next section we introduce the matching algorithm using AGROVOC and CABI.

3 Controlled Vocabulary Matching

Our problem revolves around the concept of CV matching base on the semantic matching idea described in [8]. The key intuition behind matching controlled vocabularies is the determination of mapping by computing syntactic and semantic relations which hold between the entities of any given two CVs [8, 21]. Let us consider matching 4-tuples $\langle ID_{i,j}, c_i, d_j, R\rangle$, $i = 1, ..., N_C$; $j = 1, ..., N_D$ where $ID_{i,j}$, is a unique identifier of the given mapped element; c_i is the i-th node of the CV1, N_C is number of nodes in the CV1, d_j is the j-th node of the CV2, N_D is the number of nodes in the CV2 and R specify a semantic relation which may hold between the concepts at nodes c_i and d_j. Therefore, light of the above discussion, the CV matching is defined with the following in problem: given two CV T_C and T_D compute the $N_C \times N_D$ mapped element $ID_{i,j}$, c_i, d_j, R with $c_i \in T_C$, $i = 1, ..., N_C$, $d_j \in T_D$, $= 1, ..., N_D$ and R is the strongest semantic relation holding between *concepts at node* c_i, d_j. Since we look for the $N_C \times N_D$ correspondence, the cardinality of mapping between elements can be determined to be $1 : N$. If necessary, these can also be decomposed straightforwardly into mapping elements with the 1:1 cardinality.

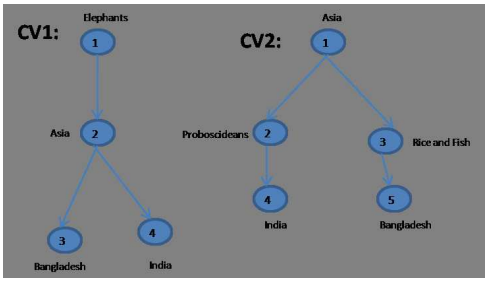

Fig. 2. Two CVs

In Figure 2, we show two CVs that contains concepts and their relations.

CV-Matching Algorithm

A Concept Facet(CF) includes combined relations, CF= $\langle lg, mg, R\rangle$, where *lg* is less general, *mg* is more general and *R* is relationships. In order to solve the

problem, we take concepts from given CVs concepts hierarchies. In our case, each concept has a concept facet. These concept facets are stored in tables for matching purpose. For instance, two concepts from node 2 "Asia" and "proboscidean" from given CVs and their concept facets appear as follows:

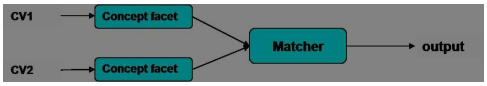

Fig. 3. CV Matching

CF (Concept Facet) of Asia :
Less general (⊑):Bangladesh,India;
More general (⊒):Elephant
CF(Concept Facet) of proboscidean:
Less general (⊑):India;
More general (⊒):Asia

Each CF contains a distinct feature for each concept. To match between two concept facets we follow the top-down approach. In our two concept facets, we start comparing with synonyms and if we find any matches between synonyms then we assume that they belongs to the same concept.Sometimes we find out similar words in different concepts (e.g. "Reading" means reading activities and also reading is a city in England). This may introduce another problem dealing with homographs.To avoid this kind of problem, we check more general and less general relationships. Furthermore, we can apply another relationship " the disjoint" which may solve the homograph problem. In addition, if concepts have the same more general term then we can say that they may be siblings.

4 Results and Evaluation: the AGROVOC and CABI case study

4.1 AGROVOC

AGROVOC is a multilingual structured and controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains (e.g. the environment). The AGROVOC Thesaurus was developed by FAO and the Commission of the European Communities in the early 1980s. Since then it has been updated continuously by FAO and local institutions in member countries. It is mainly used for indexing and retrieval data in agriculture information systems both inside and outside FAO.Its main role is to standardize the indexing process in order to make searches simpler and more efficient and to provide the user with the most relevant resources . It has approximately 20,000 concepts and four types of relations derived from

the ISO thesaurus standard: USE (a preferred term), NT (narrow terms),RT (related terms) and BT (broader terms) in XML or RDF format. We use the XML version for our task [19].

4.2 CABI

CABI is a monolingual structure controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, horticulture, soil science, entomology, mycology, parasitology, veterinary medicine, nutrition and rural studies. The CABI thesaurus was developed by CABI which is a not-for-profit, science-based development and information organization. It has 48,000 concepts and four types of relationship derived from the ISO thesaurus standard: USE (a preferred term), RT (related terms), NT (narrow terms), BT (broader terms). We obtained data as text format and converted it to XML format for experiment purposes. It is used for indexing digital or physical text, objects and collections [20].

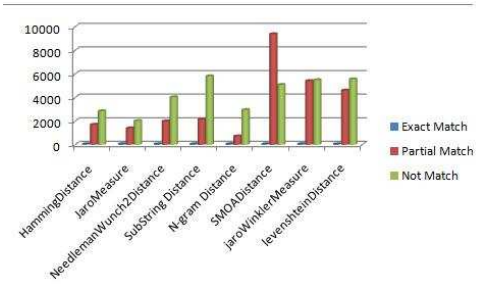

Fig. 4. cv matching results

4.3 Results and Evaluation Descriptions

In our primary experiments, we used AGROVOC thesaurus and CABI thesaurus as our controlled vocabularies because there is no complete mapping between these two thesauri. We are mapping two thesauri and would like to bring it online so that it can cover a much wider domain for indexing, searching and information retrieval purposes [13, 22]. We started our experiments using 492 concepts from each controlled vocabulary. Managing all concepts was a challenge in that two vocabularies are not organized in the same structure. We converted each vocabulary to the same format in order to conduct the test. We considered more general (Broader terms in thesaurus) and less general (Narrow terms in thesaurus) concepts from two thesauri.

We obtained 64 exact matches between terms from all element label matchers of given vocabularies but we found different levels of partial matches from eight

element label matchers. Figure 4 shows SMOADistance matcher gives more partial matches than others.Hamming distance, JaroMeasure, SubStringDistance and N-gram do not give satisfactory levels of matches. JaroWinKlerMesaure and LevesteinDistnace produce quite similar results. However, these are our primary levels of matching results for concept facets. They contains overlapping problem of terms due to same domain and existing same terms in different positions. In the future, we will evaluate our approach using structure and semantic matching techniques. We can not describe further details due to page restrictions.

5 Conclusion

In this paper, we have shown our proposed system for automatic creation of controlled vocabulary and matching two vocabularies using concept facets. We are convinced that it helps for information searching, browsing, extraction in agriculture, forestry, food, horticulture, soil science, entomology, mycology, parasitology, veterinary medicine, nutrition and rural studies fields. But, there are some open research issues already devised for designing such systems. For example, we can cite the researches on semantic heterogeneity between two controlled vocabularies in a single domain, how to reach agreement of the position of concepts in the hierarchy so that we can say they are similar and multi-word concepts. Yet, there are open questions to be answered like how to react when a certain term is not recognized, does the system rely solely on user communities to derive the concept or does it first try to search for them in external reliable sources of information such as public thesauri, encyclopedia or dictionaries.

Acknowledgment

Authors would like to thank Prof. Fausto Giunchiglia, Ilya Zaihrayeu and Vincenzo Maltese for their valuable suggestions. Also, we would like to thank Shaun Hobbs of CABI for kindly providing us with the data files.

References

1. F. Bancilhon. Naive evaluation of recursively defined relations. In M. L. Brodie and J. Mylopoulos, editors, *On Knowledge Base Management Systems*, pages 165–178. Springer, New York, 1986.
2. P. Bouquet, L. Serafini, and S. Zanobini. Semantic coordination: a new approach and an application. *In Proc. of the 2nd International Semantic Web Conference (ISWO'03). Sanibel Islands, Florida, USA*, October 2003.
3. E.McCulloch. Digital direction thesauri:practical guidance for construction. volume 54, 2005.
4. P. Shvaiko F. Giunchiglia and M. Yatskevich. Discovering missing background knowledge in onology matching. In *17th European Conference on Artificial Intelligance (ECAI 2006)*, volume 141, pages 382–386, 2006.

5. M.Marchese F.Giunchiglia and I.Zaihrayeu. Encoding classifications into lightweight ontologies. *Data Semantics VIII*, pages 57–81, 2007.
6. F.Ibekwe-SanJuan. Construction and maintaining knowledge organization tools a symbolic appraoch. volume 62, 2006.
7. F. Giunchiglia, B.Dutta, and V.Maltese. Faceted lightweight ontologies. In *LNCS*, 2009.
8. F. Giunchiglia, P. Shvaiko, and M. Yatskevich. S-match: An algorithm and an implementation of semantic matching. *In Proceedings of ESWS'04*, 2004.
9. F. Giunchiglia and I. Zaihrayeu. Lighweight ontologies. *Technical report at DIT, the University of Trento, Italy*, October 2007.
10. A.Gilchrist J.Aitchison and Bawden. Thesaurus construction and use:a practical manual. 4th ed., page 240, London, 2006. Aslib.
11. J.Euzenate and P.Shaviko. *Ontology Matching*. Springer, 1st edition, 2007.
12. KEA Automatic keyphrase extraction. http://www.nzdl.org/Kea/.
13. Sini M. Chang C. Li S. Lu W. He C. Liang, A. and J. Keizer. The mapping schema from chinese agricultural thesaurus to agrovoc. In Proceedings of the fifth Conference of the European Federation for Information Technology in Agriculture, Food and Environment and the thirdWorld Congress on Computers in Agriculture and Natural Resources, 2005.
14. Bernardo Magnini, Luciano Serafini, and Manuela Speranza. Making explicit the semantics hidden in schema models. *In: Proceedings of the Workshop on Human Language Technology for the Semantic Web and Web Services, held at ISWC-2003, Sanibel Island, Florida*, October 2003.
15. George Miller. *WordNet: An electronic Lexical Database*. MIT Press, 1998.
16. Natalya F. Noy. Semantic integration: a survey of ontology-based approaches. *SIGMOD Rec.*, 33(4):65–70, 2004.
17. LibraryofCongressClassification system. http://www.loc.gov/catdir/cpso/lcco/lcco.html/.
18. The Dewey Decimal Classification system. http://www.oclc.org/dewey/.
19. Agrovoc thesaurus. http://www.fao.org/agrovoc/.
20. CAB thesaurus. http://www.cabi.org/.
21. Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.
22. L.Finch H. Kolb W.Hage, M.Sini and G.Schreiber. The oaei food task:an analysis of a thesaurus alignment task.
23. I. Zaihrayeu, L. Sun, F. Giunchiglia, W. Pan, Q. Ju, M. Chi, and X. Huang. From web directories to ontologies: Natural language processing challenges. In *ISWC/ASWC*, 2007.