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 facets. This methodology is based on semantic matching which is different from the orthodox view of matching.

Key words: Vocabulary Mapping, Vocabulary Creation, Thesaurus, AGROVOC, CABI

1 Automatic Controlled Vocabulary Creation

Some research has been done on Controlled Vocabulary (CV) construction by automatic or semi-automatic methods [3]. These two methods can be categorized into two approaches [1]: 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 example, suppose we have two terms “Agricultural business” and “Agricultural industry”. These two terms can 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).

The two described approaches are time-consuming and need a substantial amount of human intervention. To overcome this problem, we 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 others because they use syntactic matching techniques and they do not make use of background knowledge. Because it is difficult to find the universal background knowledge, we used WordNet [7] 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 [2] 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 [5]. 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 cluster keywords according to semantic similarities. This work on automatic CV creation is still on going: we have presented the general idea and described the algorithm, but more work would need to be carried out in order to extend the testing.

2 Controlled Vocabulary Matching

A Concept Facet (CF) contains distinct features for each concept: it includes combined relations, CF= $\langle lg, mg, R\rangle$, where lg identifies less general concepts (one or more), mg identifies more general concepts (one or more) and R identifies related concepts (one or more). In order to realize a matching between two vocabularies (CV1, CV2), we consider the CF from all given CVs's concepts: for every CF of CV1, we check the matching with all CFs of CV2. These concept facets are stored in tables for matching purpose. The methodology of the matching algorithm applied to every concept, can be represented with the following picture. The matching between two concept facets follows the top-down approach and used several lexical comparison algorithms (SMOADistance, HammingDistance, JaroMeasure, SubStringDistance, N-gram, JaroWinKlerMeasure,

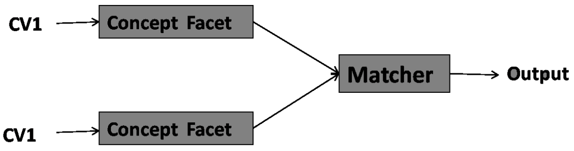

Fig. 1. CV Matching

and LavesteinDistance) [4, 8]. Firstly, we start comparing the more general concepts; if they match (they have same lexicalizations or they are synonyms) we assume that the concepts under investigation belongs to same concept (they match). Secondly (either we got match or not), we start comparing the less general concepts. Based on the results of two mentioned matching, we may obtain exact match (in case more general and less general concepts match), partial match (in case of only one match), or not match. Related concepts of CFs are considered to validate the previous results.

3 Results and Evaluation: the AGROVOC and CABI case study

In our experiments, we used the AGROVOC thesaurus and the CABI thesaurus because there is no complete mapping between them. The results of the mapping will be published online so that users can use them for better indexing, searching and information retrieval [6, 11].

3.1 AGROVOC

AGROVOC is a multilingual 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. It has approximately 20,000 concepts and four types of relations derived from the ISO standard. Among the available format, we used the XML version for our task [9].

3.2 CABI

CABI is a monolingual 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 standard. We obtained data as text format and converted it to XML format for experiment purposes [10].

3.3 Results and Evaluation Descriptions

We started our experiments using 492 concepts from each controlled vocabulary. Managing all concepts was a challenge because the two vocabularies are not organized in the same structure. We converted each vocabulary to the same format in order to conduct the test. We obtained 64 exact matches from all tested algorithms, but we found different numbers of partial matches from eight element label matchers. SMOADistance matcher gives more partial matches than others. Hamming distance, JaroMeasure, SubStringDistance, and N-gram do not give a satisfactory numbers of matches. JaroWinKlerMesaure and LevesteinDistance produce quite similar results. However, these are our primary results which should be validated by extending the process to the full thesauri.

4 Conclusion

In this paper, we have shown our proposed system for automatic creation of controlled vocabulary and vocabulary matching using concept facets. We are convinced that it helps for better information searching, browsing, and extraction in agriculture and related domains. There are some open research issues: the semantic heterogeneity between two controlled vocabularies in a single domain; the multi-word concepts; the possibility of automatically link non-matched concepts to external reliable resources 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.Ibekwe-SanJuan. Construction and maintaining knowledge organization tools a symbolic appraoch. volume 62, 2006.
2. F. Giunchiglia, P. Shvaiko, and M. Yatskevich. S-match: An algorithm and an implementation of semantic matching. *In Proceedings of ESWS'04*, 2004.
3. A.Gilchrist J.Aitchison and Bawden. Thesaurus construction and use:a practical manual. 4th ed., page 240, London, 2006. Aslib.
4. J.Euzenate and P.Shaviko. *Ontology Matching*. Springer, 1st edition, 2007.
5. KEA Automatic keyphrase extraction. http://www.nzdl.org/Kea/.
6. 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.
7. George Miller. *WordNet: An electronic Lexical Database*. MIT Press, 1998.
8. Natalya F. Noy. Semantic integration: a survey of ontology-based approaches. *SIGMOD Rec.*, 33(4):65–70, 2004.
9. Agrovoc thesaurus. http://www.fao.org/agrovoc/.
10. CAB thesaurus. http://www.cabi.org/.
11. L.Finch H. Kolb W.Hage, M.Sini and G.Schreiber. The oaei food task:an analysis of a thesaurus alignment task.