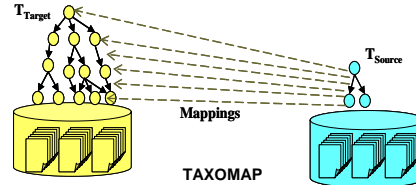


Structural techniques for alignment of structurally dissymmetric taxonomies

Chantal Reynaud, Brigitte Safar, Hassen Kefi
CNRS - Université Paris-Sud (LRI) & INRIA-Futurs (GEMO)

- Very specialized taxonomies with only sub-class links.
- Labels of concepts which are expressions composed of a lot of words.
- Words common to a lot of labels.



- One-to-one mappings.
- Equivalence or subclass relations
- An oriented mapping process from T_{Source} to T_{Target} .

General view of the alignment process

- An alignment based on Lin's similarity measure.

$$Sim_{lin}(x,y) = 2 * \frac{\sum_{t \in tr(x) \cap tr(y)} \log P(t)}{\sum_{t \in tr(x)} \log P(t) + \sum_{t \in tr(y)} \log P(t)}$$

- Various techniques applied in sequence:
 1. Terminological techniques ➤ Most likely mappings
 2. Structural/semantic techniques ➤ Interesting but less certain mappings
 - a. Exploiting the structure of T_{Target} : **STR_T**
 - b. Exploiting the structure of WordNet: **STR_W**
- Objective: For each technique, select the best concept in T_{Target} among a lot of mapping candidates.

Main contributions

- Original structural alignment techniques
 - Different from a search of structural similarity in both models
 - Usable with structural dissymmetric taxonomies.
- An efficient use of WordNet to provide an additional structural support.
- Experiments in the setting of the e.dot project on two real-world taxonomies in the field of predictive microbiology and on test taxonomies.

Exploiting the structure of the target taxonomy: STR_T

- Interesting when labels are composed of many words.

Approach

- Main idea: For each element c_s in T_{Source} , exploit the location of its mapping candidates (MC) in T_{Target} .

A lot of elements of MC have a common ancestor which is deep enough in T_{Target}

↓

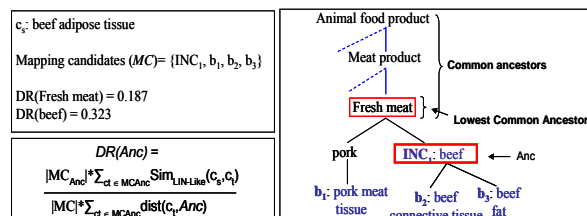
They are close elements which share a common context

↓

c_s is likely meaningful according to that context too

Design process

- Based on the computation of:
 - the Lowest Common Ancestor of the nodes of MC ,
 - Their partial ancestor nodes, Anc , and their relative density, $DR(Anc)$.



- The partial ancestor node with the highest DR is the most relevant one.
- The concept retained for the mapping with c_s belongs to the context of the most relevant Anc ; it has the highest similarity value.

Exploiting the structure of WordNet: STR_W

- Interesting to provide an additional structural support.
- Interesting to map concepts syntactically different but semantically similar without being synonyms.

Two steps

1. Build a sub-tree (a unique one), T_W , by searching Wordnet for the hypernyms of each term of T_{Target} not yet mapped.
2. Given an element c_s of T_{Source} not yet mapped, use the Wu & Palmer's measure to find the element of T_{Target} the most similar in T_W .

$$Sim_{W\&P}(c_1, c_2) = \frac{2 * \text{depth}(LCA(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}$$

Central point in the use of Wu & Palmer's measure

- The similarity is higher between c_s and any of its brothers or any of the descendants close to its brothers than between c_s and its grandfather, **until a depth p** that can be computed given any node c_s in function of its depth in the tree.

- An illustration of the search strategy

