A Similarity Measure based on Semantic, Terminological and Linguistic Information

Nitish Aggarwal*, Tobias Wunner*°, Mihael Arcan*,
Paul Buitelaar*, Seán O'Riain°

*Unit for Natural Language Processing and °eBusiness Unit
Digital Enterprise Research Institute,
National University of Ireland, Galway
firstname.lastname@deri.org

Introduction

The fundamental task of ontology matching is based on measuring the similarity between two ontology concepts [1]. However, we argue that a deeper semantic, terminological and linguistic analysis of specialized domain vocabularies is needed in order to establish a more sophisticated similarity measure that caters for the specific characteristics of this data. In particular we propose 'STL', a novel similarity measure that takes semantic, terminological and linguistic variation into account.

The STL Similarity

We base our approach on the three-faceted STL ontology enrichment process introduced in [3]. We calculate similarity according to semantic, terminological and linguistic variation and then take a linear combination by using linear regression, called STL similarity, which we describe as follows:

Semantic similarity (sim_S) is calculated based on semantic (taxonomic or ontological) structure. For our purposes we used a recently proposed semantic similarity measure proposed by Pirro_Sim [2], which uses intrinsic information content, i.e. the information content of a concept defined by the number of its subconcepts.

Terminological similarity (sim_T) is defined by maximal subterm overlap, i.e. we calculate sim_T between two concepts c1 and c2 as the number of subterms t_i in a termbase that can be matched on the labels of c1 and c2. A term t_i is said to match on a concept when no other longer term t_j can be matched on the same concept (label). To calculate sim_T we use monolingual as well as multilingual termbases as the latter reflect terminological similarities that may be available in one language but not in others, e.g. there is no terminological similarity between

the English terms "Property Plant and Equipment" and "Tangible Fixed Asset", whereas in German these concepts are actually identical on the terminological level (they both translate into "Sachanlagen").

Linguistic similarity (sim_L) is defined as the Dice coefficient applied on the head&modifier syntactical arguments of two terms, i.e., the ratio of common modifiers to all modifiers of two concepts. For instance the concepts "Financial Income" and "Net Financial Income" have 3 modifiers "financial" "net" and "net financial", whereby only "financial" is a common modifier.

Putting it all together we define STL similarity as a linear combination of the sub-measures where the weights $w_{S,T,L}$ are their contributions on the data set:

$$sim_{S,T,L} = w_S * sim_S + w_T * sim_T + w_L * sim_L + constant$$

We evaluated our approach on a data set of 59 financial term pairs, drawn from the xEBR (European Business Registry) vocabulary, that were annotated by four human annotators. Table 1 shows the correlations ρ of all measures on the data set and that STL outperforms all of its S,T and L contributions.

Measure	ρ	**Type**
PathLength	0.16	S
Wu−Palmer	0.18	S
Pirro_Sim	0.20	S
UnigramMono	0.72	T

Measure	ρ	**Type**
UnigMulti	0.72	T
BigMono	0.53	T
Bi_Multi	0.54	T
SubtermMono	0.74	T

Measure	ρ	**Type**
SubtermMulti	0.75	T
Lemmatized	0.70	L
Head&Mod	0.51	L
STL	**0.78**	S,T,L

Table 1. Correlation of STL similarity measures with human evaluator scores

References

1. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., dos Santos, C.T.: Ontology alignment evaluation initiative: Six years of experience. J. Data Semantics 15, 158–192 (2011)
2. Pirró, G.: A semantic similarity metric combining features and intrinsic information content. Data Knowl. Eng. 68, 1289–1308 (November 2009)
3. Wunner, T., Buitelaar, P., O'Riain, S.: Semantic, terminological and linguistic interpretation of xbrl. In: In Reuse and Adaptation of Ontologies and Terminologies Workshop at 17th EKAW (2010)

Acknowledgements

This work is supported in part by the European Union under Grant No. 248458 for the Monnet project and by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).