Towards a Logic-based Assessment of the compatibility of UMLS sources

E. Jiménez-Ruiz[1⋆], B.Cuenca Grau[2⋆⋆], R. Berlanga[1], and I. Horrocks[2]

[1] Universitat Jaume I, Spain, {ejimenez,berlanga}@uji.es
[2] University of Oxford, UK, {berg,ian.horrocks}@comlab.ox.ac.uk

Abstract The UMLS Metathesaurus (UMLS-Meta) is currently the most comprehensive effort for integrating independently-developed medical thesauri and ontologies. The techniques used in the construction of UMLS-Meta are mostly based on lexical matching and often disregard the semantics of the sources being integrated. In this paper we aim at developing logic-based techniques to automatically detect and fix potential errors in UMLS-Meta. Our research is currently at an early stage, so we only present here our preliminary ideas and experimental results.

1 Motivation

In its 2009AA version, UMLS-Meta [1] integrates more than one hundred thesauri and ontologies. The main content of UMLS-Meta is a list with more than two million unique identifiers (CUIs). Associated to each CUI, there is a set of term names coming from different sources. Pairs of terms with the same CUI are synonyms and hence can be represented as an equivalence mapping.

Currently, the integration of new sources in UMLS-Meta combines automatic techniques together with expert assessment [1]. Automatic techniques are mainly based on lexical matching algorithms (e.g., [2]). Other techniques used to improve the design process involve, for example, exploiting synonymy relations from external knowledge sources such as WordNet (e.g., [3]).

The main limitation of these techniques is that they do not take into account the logic-based semantics of the sources, which can be rich ontologies, rather than simple taxonomies (e.g., FMA, NCI, and SNOMED). Our ultimate goal is to develop logic-based techniques to detect both potential errors and missing information in both UMLS-Meta and such rich ontologies. Our preliminary results using heuristics inspired in logic-based reasoning and module extraction suggest, on the one hand, that UMLS-Meta might be incomplete and, on the other hand, that it contains a fair number of conflicting mappings, which reveal potential design errors in either UMLS-Meta and/or in the integrated ontologies. We also propose novel techniques for automating the conflict disambiguation process.

⋆ Ernesto Jimenez was supported by the Valencian Government (BFPI06/372).
⋆⋆ Bernardo Cuenca is supported by a Royal Society University Research Fellowship.

2 Proposed Principles

The logic-based techniques we aim at developing are based on the three general principles that we propose next.

> **Conservativity Principle**: The mappings alone should not introduce new semantic relationships between concepts from one of the sources.

For example, UMLS-Meta contains two mappings establishing the equivalence between the concept *Cardiac_Muscle_Tissue* from FMA and the NCI concepts *Myocardium* and *Heart_Muscle* respectively. As a consequence, UMLS-Meta implies that *Myocardium* is also equivalent to *Heart_Muscle*. However, in NCI *Myocardium* neither subsumes, nor it is subsumed by *Heart_Muscle*. The conservativity principle suggests that the obtained mappings are in conflict and (at least) one of them may be incorrect.

> **Consistency Principle**: The integration of well-established ontologies should not introduce unintended logical consequences.

For example, UMLS-Meta maps the FMA concept *Protein* to the NCI concept *Protein*, and the FMA concept *Lymphokine* to the NCI concept *Therapeutic_Lymphokine*. In FMA, *Lymphokine* is a type of *Protein*, whereas in NCI *Therapeutic_Limphokine* is a type of *Drug*. Furthermore, *Drug* and *Protein* are disjoint in NCI and hence the union of NCI, FMA and UMLS-Meta would imply that *Lymphokine* and *Therapeutic_Limphokine* are unsatisfiable.

Inconsistencies and other unintended logical consequences may be due to either erroneous mappings or to inherent incompatibilities between the sources. In any case, if the integrated sources are to be successfully used in an application, these errors should be fixed by modifying either the sources or the mappings.

> **Locality Principle**: If two concepts C and C' from ontologies $\mathcal{O}$ and $\mathcal{O}'$ are correctly mapped, then the concepts semantically related to C in $\mathcal{O}$ are likely to be mapped to those semantically related to C' in $\mathcal{O}'$.

If the locality principle does not hold, then UMLS-Meta may be incomplete and new mappings should be discovered, or the definitions of both concepts in their respective ontologies may be different or incompatible, or the mapping between C and C' may be erroneous. As an example of the latter, UMLS-Meta maps the concepts *Upper_Extremity* from NCI and *Arm* from FMA. The mapping violates the locality principle because none of the entities in their respective logic-based modules [4] have been mapped. After closer inspection of the ontologies, the mapping can be clearly identified as erroneous.

3 Implemented Heuristics

To implement these principles, we propose a preliminary collection of heuristics. The first two heuristics given next are related to similar ones used by [5, 6, 7] in a different setting. The third one is, to the best of our knowledge, entirely novel.

Injectivity of mappings. If concepts C_1 and C_2 from $\mathcal{O}$ are mapped via UMLS-Meta to the same concept D from $\mathcal{O}'$, then UMLS-Meta alone implies that C_1 and C_2 are logically equivalent. However, if $\mathcal{O}$ does not imply the equivalence of C_1 and C_2 then the conservativity principle is violated (see previous example). In that case, we say that these mappings are in conflict.

Disjointness-based inconsistency. If C_1 and C_2 from $\mathcal{O}$ are mapped to D_1 and D_2 from $\mathcal{O}'$ and $\mathcal{O}$ implies that C_1 is subsumed by C_2, but $\mathcal{O}'$ implies that D_1 and D_2 are disjoint, then the consistency principle is violated (see previous example). A variant of this heuristic, which we call *assumption of disjointness*, is obtained by recording a conflict whenever no subsumption relationship holds between D_1 and D_2 (and not only if they are disjoint). This reflects the fact that ontologies are typically underspecified w.r.t. disjointness.

Similarity of logic-based modules. To formalize the notion of a concept being "semantically related" to another concept in an ontology, we use the well-known modularization framework from [4]. If C from $\mathcal{O}$ is mapped via UMLS-Meta to D from $\mathcal{O}'$, and most of the concepts in the *module* $M_C^{\mathcal{O}}$ for C in $\mathcal{O}$ are not mapped to those in the module $M_D^{\mathcal{O}'}$ for D in $\mathcal{O}'$, then the locality principle is violated (see previous example). In this case the mapping between C and D is recorded as "suspicious". To implement this idea, we measure the similarity between the corresponding modules by computing the relationship between the number of concepts in the modules which are mapped via UMLS-Meta and those which are not, using an adaptation of the well-known Dice's coefficient:

$$\mathsf{sim}(M_C^{\mathcal{O}}, M_D^{\mathcal{O}'}) = 2 \times \frac{\mid \text{ Mappings between } \mathsf{sig}(M_C^{\mathcal{O}}) \ \& \ \mathsf{sig}(M_D^{\mathcal{O}'}) \mid}{\mid \mathsf{sig}(M_C^{\mathcal{O}}) \mid + \mid \mathsf{sig}(M_D^{\mathcal{O}'}) \mid} \qquad (1)$$

where $\mathsf{sig}(\cdot)$ denotes the set of concepts and relationships in the corresponding module. If the similarity between the modules of the mapped entities is lower than a given threshold, we assume that the mapping is "suspicious".

The first two heuristics allow us to identify pairs of mappings in UMLS-Meta that are (potentially) in mutual conflict. However, it is not clear how to automatically disambiguate these conflicts. To this end, we again exploit the locality principle. Assume that C_1 and C_2 from $\mathcal{O}$ are mapped via UMLS-Meta to D_1 and D_2 from $\mathcal{O}'$ respectively and that these mappings are in conflict. We then compute the similarities $\mathsf{sim}(M_{C_1}^{\mathcal{O}}, M_{D_1}^{\mathcal{O}'})$ and $\mathsf{sim}(M_{C_2}^{\mathcal{O}}, M_{D_2}^{\mathcal{O}'})$ respectively as in (1) and select the mapping with the highest associated similarity.

4 Preliminary Experiments and Future Work

We have evaluated our heuristics using UMLS-Meta version 2009AA and the corresponding versions of FMA, SNOMED and NCI. FMA, NCI and SNOMED contain 78989, 66724 and 304802 concepts respectively. UMLS-Meta 2009AA contains 2271 mappings between FMA and NCI, 8376 mappings between FMA and SNOMED and 18384 mappings between SNOMED and NCI.

Using the principle of conservativity, we have found 513 conflicting pairs of mappings between FMA and NCI, 1367 between FMA and SNOMED and 4290 between SNOMED and NCI. Using logic-based modules as explained in the end of Section 3, we obtained that 239 mapping pairs between FMA and NCI (resp. 65 between FMA and SNOMED, and 1158 between SNOMED and NCI) could not be disambiguated since no other concepts in the relevant modules where mapped by UMLS-Meta. For the remaining pairs we could produce a recommendation.

To evaluate the principle of consistency, we concentrate on the mappings between NCI and FMA:

- Using the *disjointness-based inconsistency* heuristic we found 307 conflicting mapping pairs between FMA and NCI. Using logic-based modules, we failed to disambiguate only 36 conflicting pairs. Each of these conflicts will *certainly* lead to the unsatisfiability of a concept in the union of the source ontologies and UMLS-Meta. Thus, semantically, the integration of these ontologies via UMLS-Meta is far from error-free.
- Using the *assumption of disjointness* heuristic we found 1707 conflicts between FMA and NCI. We failed to disambiguate only 202 conflicting pairs.

Finally, using the principle of locality and a similarity threshold of 1% (resp. 2%) we could identify 12 (resp. 110) "suspicious" mappings between FMA and NCI, 10 (resp. 689) between FMA and SNOMED and 1420 (resp. 2336) between SNOMED and NCI. This implies that there is a significant number of mappings whose "semantic neighborhood" is not mapped accordingly.

Previous results suggest the benefits of the implemented heuristics in the design of normative mapping sets such as UMLS-Meta. For future work, we plan to design new heuristics using the general principles from Section 2 and seek feedback from domain experts in the conflict disambiguation process.

References

[1] Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research **32**(Database issue) (January 2004)

[2] Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. Proc AMIA Symp (2001) 17–21

[3] Huang, K.C., Geller, J., Halper, M., Perl, Y., Xu, J.: Using wordnet synonym substitution to enhance umls source integration. Artif. Intell. Med. **46**(2) (2009)

[4] Cuenca Grau, B., Horrocks, I., Kazakov, Y., Sattler, U.: Just the right amount: Extracting modules from ontologies. In: Proc. of WWW 2007. (2007) 717–727

[5] Meilicke, C., Stuckenschmidt, H., Tamilin, A.: Reasoning Support for Mapping Revision. Journal of Logic and Computation (2008)

[6] Jean-Mary, Y.R., Shironoshita, E.P., Kabuka, M.R.: Ontology matching with semantic verification. Journal of Web Semantics (2009)

[7] Jimenez-Ruiz, E., Cuenca Grau, B., Horrocks, I., Berlanga, R.: Ontology integration using mappings: Towards getting the right logical consequences. In: Proc. of ESWC. (2009) 173–187