

Concept Mismatches Between a Source and Target Natural Language

Frances Gillis-Webber¹

¹Computer Science Department, University of Cape Town, Cape Town, South Africa

Abstract

Numerous mismatches have been identified when aligning heterogeneous resources. In this paper, the focus is on the mismatches for a concept between a source and target viewpoint, where each viewpoint is natural language-specific. A concept is first defined as a 6-tuple, comprising of its viewpoint, the lexical realisation of the concept, the axiomatisation thereof, as well as asserted individuals. The same concept is then defined as another tuple, this time for a target viewpoint, with each element therein compared to the original. A total of 22 mismatches and correspondences have been identified, with three pertaining to lexical realisations, twelve pertaining to the axiomatisation of a concept, and seven pertaining to individuals and assertions.

Keywords

ontology matching, multilingualism, ontology localisation

1. Introduction

The approaches for modelling a multilingual ontology include the use of multilingual labels, or the use of a linguistic model to associate multilingual information with ontology entities. As an alternative approach to render an ontology multilingual, a crosslingual ontology mapping method can be used to align two different natural language resources. Examples of mapping methods include the alignment of two monolingual ontologies or vocabularies using OWL's `EquivalentClass`, or SKOS for aligning individuals. Alternatively, WordNet or a linguistic model such as OntoLex-Lemon can be used as an interlingua when mapping the class labels of one or more ontologies. For each of these mapping methods, alignment is typically between heterogeneous data sources, where examples of heterogeneity include differing knowledge representation formalisms, and expressivity [1, 2].

The internationalisation goal of OWL was to support “the development of multilingual ontologies, and potentially provide different views of ontologies that are appropriate for different cultures” [3]. Within the context of natural language, a viewpoint can be defined as the view or perspective of a community, where this community is unified by some natural language and this language is an embodiment of that community's culture [4]. In this paper, the possible correspondences and mismatches are identified for a concept, when compared between a source


2nd Workshop on Modular Knowledge, 9th Joint Ontology Workshops (JOWO 2023), co-located with FOIS 2023, 19-20 July, 2023, Sherbrooke, Québec, Canada

✉ fgilliswebber@cs.uct.ac.za (F. Gillis-Webber)

🆔 0000-0002-3740-5904 (F. Gillis-Webber)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

and target viewpoint. The resources for both the source and the target are assumed to be homogenous, where examples of homogeneity include the same modelling style, the same foundational ontology used, and the same OWL expressivity. The result is that twenty-two mismatches and correspondences have been identified, in contrast to the eight mismatches that have been identified by Visser et al. [1].

The remainder of the paper is structured as follows. A concept is first defined in Section 2, followed by the identified (mis)matches between selected concept elements. There is a brief discussion in Section 3, and the paper concludes with Section 4.

2. Concept (Mis)matches Between a Source and Target Viewpoint

For a concept Co , a natural language may divide up the ‘space’ of Co differently to that of another natural language. A well-known example in the literature is the concept lexicalised in English as ‘river’, which has a natural language description of “a flowing natural watercourse”. In French, the same concept is lexicalised as ‘rivière’ and ‘fleuve’, where the former refers to rivers that flow into another river, and the latter refers to rivers that flow into the sea, and both are natural watercourses.

Expanding on the notion of the ‘space’ of a concept, we define a concept space as a 6-tuple $CS = \langle Co, VP, LI, SC, AP, Ind \rangle$, where Co is the concept, represented as a natural language description in a similar vein to C in Visser et al.[1]. VP is the viewpoint, LI is the lexical item (or label), SC is the superclass, AP is the axiom pattern, and Ind is the set of individual assertions. AP is a 2-tuple $\langle APC, Ax \rangle$ where APC is the set of axiom pattern classnames, and Ax is the set of axioms pertaining to the ontological commitment of each element in APC . Each element in APC is subsumed by SC . When comparing a concept between a source and target viewpoint, a source and target concept space is paired. We define a paired concept space as a 3-tuple $PCS = \langle CS, CS', PVP \rangle$ where CS and CS' is a concept space, and $CS \neq CS'$. PVP is the paired viewpoint within which CS and CS' is considered. A PCS is visualised in Figure 1, for the concept space of ‘river’ shown for three viewpoints: English, Afrikaans, and French.

Selected elements from CS and CS' can be compared for equivalence. However, before doing so, it is assumed that CS and CS' is homogenous, where all unary and binary predicates are normalised for case and tense (if using descriptive fragment identifiers), and all axioms are normalised for the same Description Logic naming schema and OWL serialisation. Starting first with LI and LI' from CS and CS' respectively, the identified (mis)matches are enumerated below, followed by language examples.

- M1: Correspondence with lexical gap:** there is no lexical realisation for $LexItem$ or $LexItem'$, or both.
- M2: Lexically realised correspondence:** there is a lexical realisation for both $LexItem$ and $LexItem'$.
- M3: Lexically realised correspondence with grammatical inequivalence:** there is a lexical realisation for both but there is a grammatical inequivalence between the two, in that the language feature used by the one natural language is not used by the other.

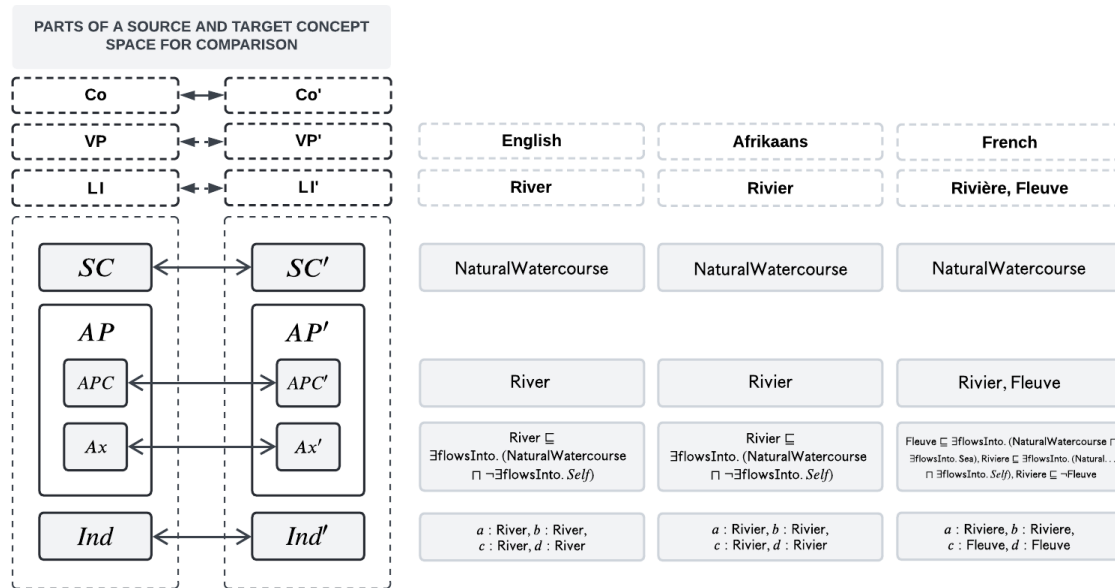


Figure 1: The parts of a source and target concept space CS and CS' . SC is the superclass, AP is the axiom pattern, consisting of a set of axiom pattern class names, APC , and axioms Ax for the ontological commitment of APC . Ind is the set of individuals asserted for each element in APC . The concept space for the concept 'river' is given for three viewpoints: English, Afrikaans, and French.

For M1, an example is the concept lexicalised in English as 'adoption' (of a child), where in the South African language, Sesotho, there is no lexical item, and instead, a paraphrase of the English term is used. The lexical item 'fleuve' from French is another example of a lexical gap in English, where there is no direct equivalent term. In these scenarios, CS' , as the target concept space for which there is a lexical gap, is then a translation of CS , with CS' assuming the axioms of CS . For M3, an example of a language feature is grammatical gender, where a lexical item is modified according to the gender of the subject. The lexical item 'priest' is one such example, where in isiXhosa, an Nguni language in South Africa, the masculine is 'umfundisi' and the feminine 'umfundisikazi'.

When comparing superclass SC to superclass SC' from CS and CS' respectively, the identified (mis)matches are as follows:

- M4: Equivalent superclasses:** both SC and SC' is equivalent. There is no mismatch.
- M5: Empty superclass for source or target:** SC or SC' is empty (and the corresponding axiom pattern is a sub-class of $\text{owl} : \text{Thing}$).
- M6: Empty superclass for source and target:** SC and SC' are empty, and both AP and AP' are a sub-class of $\text{owl} : \text{Thing}$.
- M7: Direct superclass is equivalent to indirect superclass:** the direct superclass for AP and AP' is not equivalent, however if the class hierarchy of the source or target is traversed recursively, at some point, there is a shared class between SC and SC' .
- M8: No shared direct or indirect superclass:** the direct superclass is not the same for both SC and SC' , nor is there a shared indirect class.

For M7, an example is the concept lexicalised in English as ‘spoon’. The same concept is lexicalised in Afrikaans as ‘lepel’, except that a spoon is a ‘utensil’ (which is in turn a ‘tool’), whereas in Afrikaans, a ‘lepel’ is a ‘gereedskap’ (translated as ‘tool’). The direct superclass of each is not equivalent, however, if the class hierarchy of ‘spoon’ is recursively traversed, then the indirect superclass ‘tool’ is equivalent to ‘gereedskap’. Moving on to M8, this pertains to the classification of the superclass in a class hierarchy, where SC and SC' are located in different branches of the class hierarchy. An example is the English term ‘traditional healer’, which is defined as a complementary or alternative health practitioner specialising in traditional African medicine, whereas the isiXhosa equivalent ‘igqirha’, can be defined as a doctor specialising in traditional medicine (with a ‘medical doctor’ similarly defined in isiXhosa as a doctor specialising in biomedicine).

When comparing a source axiom pattern AP to a target axiom pattern AP' from CS and CS' respectively, then the identified (mis)matches are as follows:

M9: Equivalent classes and equivalent ontological commitment: APC and APC' are equal, with the axioms of Ax and Ax' also equivalent. There is no mismatch.

M10: Equivalent classes with no ontological commitment: APC and APC' are equal, but there are no axioms expressing the ontological commitment for them both. There is no mismatch.

M11: Equivalent classes only: only APC and APC' are equal.

M12: Some shared classes: only some of the classes in APC and APC' are shared.

M13: No shared classes: there are no shared classes in APC and APC' .

M14: Ontological commitment mismatch: the axioms in Ax and Ax' are not equivalent.

M15: Reification mismatch: a refinement of M14, this is a mismatch between Ax and Ax' , where implicit knowledge is reified in the one but not the other.

For M12, an example is that of human settlements, classified in English by ‘city’, ‘town’, ‘village’, and ‘hamlet’, with the equivalent in French being ‘ville’, ‘village’, ‘bourg’, ‘bourgade’, and ‘hameau’. The only shared classes between the two languages is that for ‘city’ and ‘ville’, and for ‘hamlet’ and ‘hameau’. For M13, the English ‘river’, and French ‘rivière’ and ‘fleuve’ is an example where there are no shared classes. For M14, the isiXhosa term ‘ikhazi’ is used as a translation equivalent for English’s ‘dowry’, where money and other goods and property are brought by a bride into her marriage. However, when considered from the perspective of the AmaXhosa (the first language speakers of isiXhosa), there is a meaning distinction for ‘ikhazi’, where cattle or money is paid by the future groom as part of the bride price. For M15, an example is the English ‘pasta’ to the Italian ‘pasta’, where the English term is a borrowing from Italian for which there has been no morphemic modification. The same word is used in both languages, but the definiens for English would include an axiom to indicate that this is Italian cuisine. From the Italian viewpoint, the fact that pasta is part of their cuisine is implicit knowledge that is not likely to be made explicit.

We finish with the set of individuals Ind and Ind' from CS and CS' respectively, for which the identified (mis)matches are as follows:

- M16: No correspondence in all interpretations:** there are no shared individuals between CS and CS' . This means there is also no mismatch.
- M17: No correspondence in some interpretations:** there is a correspondence in some interpretations for CS and CS' .
- M18: Same individuals with equivalent assertions:** the same individuals are shared between CS and CS' , both with the same assertions between the two.
- M19: Same individuals but differing assertions:** the same individuals are shared between CS and CS' , but the assertions between both do not correspond. This is an example of granularity mismatch.
- M20: (Proper) subset of shared individuals:** all of the individuals of CS are a proper subset of the individuals of CS' (or vice versa). This is another example of granularity mismatch.
- M21: Some shared individuals:** this is an intersection of the elements in Ind and Ind' . This is an example of overlapping meaning.
- M22: No shared individuals but there is conceptual equivalence:** no individuals are asserted for CS and CS' , however Co and Co' can be compared for equivalence in our minds. If there is conceptual equivalence, then this is a correspondence.

For M16, if there is a paired concept space where the concept for CS has the lexical item 'dog' and the concept for CS' has the lexical item 'house', there would be no correspondence. For M17, the 'dowry' and 'ikhazi' example applies here. From an English viewpoint, there is a correspondence between CS and CS' . However, from an isiXhosa viewpoint, there is no correspondence. For M18, the example of English 'river' to Afrikaans 'rivier' applies here, where the individuals are equal between the two, as well as the assertions for each class in APC and APC' (as shown in Figure 1). For M19, the example of English 'river' to French's 'rivière' and 'fleuve' is an example of equal individuals but differing assertions. For M20, an example is English's 'electricity' to 'ugesi', the isiZulu equivalent, where isiZulu is another Nguni language local to South Africa. In isiZulu, 'ugesi' originally meant 'gas', but the meaning has since extended to include 'electricity' as well. For M21, the example of 'traditional healer' and 'igqirha' applies here. Lastly, for M22, this correspondence applies where there are no shared individuals in an ontology between both concepts, or there may be no individuals asserted at all in the ontology.

3. Discussion

Of the mismatches identified by Visser et al. [1], Klein [2], and Euzenat and Shvaiko [5], these primarily pertain to the conceptualisation of the domain before formalisation, the selected representation language, and modelling decisions, such as the choice of foundational ontology. For logic language-dependent mismatches, a terminological mismatch was identified by each, with Visser et al. identifying five additional mismatches regarding the definition of a term for the elements Co , LI , and a combined SC and AP as the definiens [1].

In this paper, correspondences and mismatches have been detailed for a source and target viewpoint, where the focus has been on the term, the subsequent vocabulary and ontological

commitment in the ontology, and the asserted individuals. The transformation process from a source to a target viewpoint is current work. Instead of maintaining a separate ontology for each viewpoint, each concept that differs in the target viewpoint to that in the source ontology is modelled as a small ontology, using the same modelling style and vocabulary as that of the source ontology. An RDF file is also created to identify the mismatch types, metadata, and any refactoring axioms. Using an algorithm, each of the modules specific to the target viewpoint are imported into the source ontology, and the source ontology is refactored according to the mismatches identified. The focus has been on a language-specific viewpoint, however, mismatches can also be identified for another viewpoint (*PVP* from a paired concept space), such as a pivot natural language (for those paired concept spaces for which there is neither a source nor target lexical realisation, so another natural language is used as the translation).

4. Conclusion

By representing a concept as a tuple for each viewpoint, the elements in both the source and target tuple could then be compared in more detail. The result is that 22 correspondences and mismatches were identified, three for lexical realisations, twelve pertaining to the axiomatisation of a concept, and the remaining seven pertaining to individuals and assertions.

Acknowledgments

This work was financially supported by Hasso Plattner Institute for Digital Engineering through the HPI Research School at UCT.

References

- [1] P. R. Visser, D. M. Jones, T. Bench-Capon, M. Shave, Assessing Heterogeneity by Classifying Ontology Mismatches, in: N. Guarino (Ed.), *Formal Ontology in Information Systems: Proceedings of the First International Conference (FOIS'98)*, June 6-8, 1998, Trento, Italy, IOS Press, 1998, pp. 148–162.
- [2] M. Klein, Combining and Relating Ontologies: An Analysis of Problems and Solutions, in: A. Gomez-Perez, M. Gruninger, H. Stuckenschmidt, M. Uschold (Eds.), *Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing*, Seattle, USA, August 4-5, 2001, volume 47 of *CEUR Workshop Proceedings*, CEUR-WS, 2001, pp. 53–62.
- [3] W3C OWL Working Group, *OWL Web Ontology Language Use Cases and Requirements: W3C Recommendation 10 February 2004*, W3C Recommendation, World Wide Web Consortium, 2004. URL: <https://www.w3.org/TR/webont-req/>, online; accessed: 2023, April 28.
- [4] F. Gillis-Webber, Towards an Ontology of Viewpoints, in: *Proceedings of the 13th International Conference on Formal Ontology in Information Systems (FOIS 2023)*, 17–20 July, Sherbrooke, Québec, Canada, 2023.
- [5] J. Euzenat, P. Shvaiko, *Ontology Matching: Second Edition*, Springer-Verlag Berlin Heidelberg, 2013. doi:10.1007/978-3-642-38720-3.