Context Discovery via Theory Interpretation

Oliver Kutz[1] and **Immanuel Normann**[2]

[1]Research Center on Spatial Cognition (SFB/TR 8), University of Bremen, Germany

[2]Department of Linguistics and Literature, University of Bremen, Germany

okutz@informatik.uni-bremen.de, normann@uni-bremen.de

Abstract

We report on ongoing work to apply techniques of automated theory morphism search to ontology matching and alignment problems. Such techniques are able to discover 'structural similarities' across different ontologies by providing theory interpretations of one ontology into another. In particular, we discuss two such scenarios: one where the signatures and logics of the component ontologies fit enough to directly translate one ontology into the other, called *stringent contexts*, and one where we need to lift the ontologies to first-order logic, possibly extended by definitonal axioms introducing extra non-logical symbols, called *conforming contexts*. We also sketch the techniques currently available for automating the task of finding theory interpretations in first-order logic and discuss possible extensions.

1 Introduction and Motivation

The problem of finding semantically well-founded correspondences between ontologies, possibly formulated in different logical languages, is a pressing and challenging problem. Ontologies may be about the same domain of interest, but may use different terms; one ontology might go into greater detail than another, or they might be formulated in different logics, whilst mostly formalising the same conceptualisation of a domain, etc. To allow re-use of existing ontologies and to find overlapping 'content', we need means of identifying these 'overlapping parts'.

Often, ontologies are mediated on an ad-hoc basis. Clearly, any approach relying exclusively on lexical heuristics or manual alignment is too error prone and unreliable, or does not scale. As noted for instance by [Meilicke *et al.*, 2008], even if a first matching is realised automatically using heuristics, a manual revision of such candidate alignments is still rather difficult as the semantics of the ontologies generally interacts with the semantics given to alignment mappings.

A new approach, that we currently explore, is to apply methods of automated theory interpretation search to the realm of ontologies. Such methods have been mainly developed for the application to formalised mathematics, and some of the techniques are specialised for theories formulated in first-order logic. Theory interpretations have a long history in mathematics generally, and are probably employed by any 'working mathematician' on a daily basis; the basic idea is the following: given two theories T_1 and T_2, find a mapping of terms of T_1 to terms of T_2 (a signature morphism, typically expected to respect typing) such that all translations of axioms of T_1 become provable from T_2. If such a theory interpretation is successfully provided, all the knowledge that has already been collected w.r.t. T_1 can be re-used from the perspective of T_2, using the translation (see [Farmer, 1994] for some examples from the history of mathematics). In this case, in mathematical jargon, we might say that T_2 **carries the structure** of T_1.

An abundance of notions of context have been studied in the literature (see e.g. [Serafini and Bouquet, 2004]). We here propose to use a notion of context, more precisely contextual interpretation of an ontology, inspired by the notion of theory interpretation from mathematics, which in practice is used as a *structuring device* for mathematical theories.

Certain, very basic structures, are found everywhere in mathematics. The most obvious example might be group theory. The basic abstract structure of a group can be re-interpreted in a more concrete setting, giving the group in question additional structure (think of the natural numbers, rings, vector-spaces, etc.). Re-using the metaphor mentioned above, we say that an ontology O_2 *carries the structure of* O_1, if the latter can be re-interpreted, by an appropriate translation σ, into the language of O_2 such that all translations of its axioms are entailed by O_2. In this case, informally, we consider the pair $\langle O_2, \sigma\rangle$ a **context** for O_1.

2 Ontology Interpretations and Context

For simplicity and lack of space, we here focus on the case of ontologies formulated in first-order logic (**FOL**) or standard description logics (DLs), and omit some of the technical details. More precisely, we limit the investigation here to the case of **FOL** and to DLs that have a *standard translation* into first-order logic (which of course are conservative).

In this setting, given an ontology O, i.e. in the case of DL a set of Abox, Tbox, and Rbox statements, the **signature** of O, denoted $\mathbf{Sig}(O)$ is the set of *non-logical symbols*, i.e. object, concept, and role names found in O, i.e. the set of nullary (constants), unary, and binary (or in general n-ary) predicates, when seen from the first-order perspective. Given two ontologies O_1, O_2, an **ontology signature morphism** (mop for short) is any map $\sigma : \mathbf{Sig}(O_1) \to \mathbf{Sig}(O_2)$ respecting typing, i.e. mapping concept names to concept names, role names to role names, and object names to object names. If such a σ exists, we call the signatures **fitting**, written $O_1 \boxdot\!\!\to O_2$.

By the **logic** of O, written $\mathcal{L}(O)$, we mean the set of *logical symbols* used in O (and thus provided by the underlying DL or **FOL**), and by the **language** of O we mean the set of all well-formed formulae that can be build from the signature $\mathbf{Sig}(O)$ using the logic $\mathcal{L}(O)$.

We distinguish between directly and indirectly interpretable ontologies. An ontology O_1 is **directly interpretable** into an ontology O_2 if $\mathcal{L}(O_1) \subseteq \mathcal{L}(O_2)$ (i.e. the set of logical symbols used in O_1 are a subset of those used in O_2),[1] written $O_1 \odot\!\!\rightarrow O_2$, and $\mathbf{Sig}(O_1), \mathbf{Sig}(O_2)$ are fitting, i.e. $O_1 \boxdot\!\!\rightarrow O_2$. Otherwise, we call them **indirectly interpretable**. To illustrate this concept, if for instance we have $\mathcal{L}(O_1) = \{\forall^{\mathrm{DL}}, \sqcap\}$ just using universal restrictions and conjunctions (where the underlying DL of O_1 is $\mathcal{ALC}$), and $\mathcal{L}(O_2) = \{\exists^{\mathbf{FOL}}, \vee\}$ just using existential quantification and disjunction (where **FOL** is the underlying logic), O_1 is only indirectly interpretable in O_2, because, although the latter is of course strictly more expressive, it requires a definition of the logical operators of $\mathcal{ALC}$ within **FOL**, accomplished via the standard translation into **FOL**.

The significance of these distinctions can be seen from:

Definition 1 (Canonical Sentence Translation) *Let O_1, O_2 be ontologies, and assume O_1 is directly interpretable into O_2. Then every* mop *σ will, by a straightforward structural induction over the grammar of that DL (or* **FOL***), yield a* ***sentence translation*** *$\hat{\sigma}$ of the axioms of O_1 along σ into the language of O_2.*

However, whenever either the logics or the signatures of the ontologies involved do not directly fit, there are a number of possible solutions to choose from (we can just extend the logic in question, we can extend definitionally the signature, or both).[2] We here provide a uniform solution as follows:

Definition 2 (Derived Sentence Translation) *Suppose O_1 is only indirectly interpretable in O_2. Let λ_i denote the standard translation from $\mathcal{L}(O_i)$ into* **FOL***, $O_i' = \lambda_i(O_i)$, and let $\mathcal{S} \supseteq \mathbf{Sig}(O_2')$ such that $\mathbf{Sig}(O_1') \boxdot\!\!\rightarrow \mathcal{S}$ for a signature morphism $\widetilde{\sigma}$ such that $\widetilde{\sigma}|_{\mathbf{Sig}(O_2')} = \mathsf{id}$ (this always exists). Let $\widetilde{O_2'}$ result from O_2' by adding, for each element of $\mathcal{S} \setminus \mathbf{Sig}(O_2')$ a definitional axiom in the language of* **FOL***. By construction, $O_1' \odot\!\!\rightarrow \widetilde{O_2'}$. Now, define the* ***derived sentence translation*** *$\widetilde{\sigma}$ as the canonical sentence translation map in* **FOL***, induced by $\widetilde{\sigma}$.*

The situation in Def. 2 is illustrated in Fig. 1. We can now define the notion of *ontology interpretation*:

Definition 3 (Stringent Interpretations and Context) *Let O_1, O_2 be ontologies, suppose O_1 is directly interpretable into O_2, and let $\sigma : \mathbf{Sig}(O_1) \to \mathbf{Sig}(O_2)$ be a* mop*.*

$\sigma : O_1 \to O_2$ is called a ***stringent ontology interpretation*** *(*sop*) if $O_2 \models \hat{\sigma}(O_1)$. In this case, $\langle \sigma, O_2 \rangle$ is called a* ***stringent context*** *for O_1.*

Stringent interpretations cover the case where we can 'embed', within the same logic, one ontology into another, thus 'strictly aligning' the resp. terminologies. Let us next look at the more complex heterogeneous case of only indirectly interpretable ontologies.

Definition 4 (Conforming Interpretations and Context) *Let O_1, O_2 be ontologies, and suppose O_1 is only indirectly interpretable into O_2. Let σ be a maximally partial signature morphism, and assume $\widetilde{\sigma} : \mathbf{Sig}(O_1') \to \mathbf{Sig}(\widetilde{O_2'})$ be a* mop *in* **FOL***, where $\mathfrak{S}$ is the set of additional definitional axioms.*

$\widetilde{\sigma} : O_1 \to O_2$ is called a ***conforming ontology interpretation*** *(*cop*) if $\widetilde{O_2'} \models \widetilde{\sigma}(O_1')$. In this case, $\langle \widetilde{\sigma}, \mathfrak{S}, O_2 \rangle$ is called an O_1-**conforming context**.*

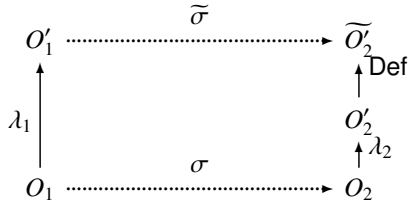

Figure 1: Ontology Interpretations

Note that the notion of a conforming context is closely related to the *heterogeneous refinements* defined in [Kutz *et al.*, 2008]: namely, given any O_1-conforming context $\langle \widetilde{\sigma}, \mathfrak{S}, O_2 \rangle$, O_2 is a heterogeneous refinement of O_1 where the transition from O_2 to $\widetilde{O_2'}$ is not only conservative but in fact definitonal.

Here is an illustrative example from mathematics:

Example 5 (Lattices and Partial Orders) *Consider P as the theory of partial-orders with $\mathbf{Sig}(P) = \{\leq\}$ and let L be the theory of lattices with $\mathbf{Sig}(L) = \{\sqcap, \sqcup\}$. These are both first-order theories, so we have $P \odot\!\!\rightarrow L$. However, the signatures obviously do not fit as L has no binary relations, so we have $P \not\boxdot\!\!\rightarrow L$. Extend the signature of L by a binary relation symbol $\sqsubseteq$ (which makes the signatures fit by the mapping $\widetilde{\sigma} : \leq \mapsto \sqsubseteq$), and define $\mathfrak{S} = \{\forall a, b. a \sqsubseteq b \leftrightarrow a \sqcup b = a\}$. This is a definitional axiom. It can now be seen that $L \cup \mathfrak{S} \models \widetilde{\sigma}(P)$, i.e. $\widetilde{\sigma}$ is a* cop*, and thus $\langle \widetilde{\sigma}, \mathfrak{S}, L \rangle$ is a P-conforming context.*

Thus, we may say that lattices carry the structure of partial orders. It should be obvious that both these theories also define central structures for ontology design.

3 Automated Discovery of Contexts

The goal of discovering ontology interpretations may be rephrased as the problem of finding all those ontologies in a large repository $\mathfrak{R}$ that could serve as a (stringent or conforming) context for a given ontology O_1. More formally, given O_1, we are looking for the sets

$$\{O_2 \in \mathfrak{R} \quad | \quad \langle \sigma, O_2 \rangle \text{ stringent context for } O_1\}$$

and

$$\{O_2 \in \mathfrak{R} \quad | \quad \langle \widetilde{\sigma}, \mathfrak{S}, O_2 \rangle \text{ conforming context for } O_1\}$$

In case of ontologies formalised in **FOL**, this task is undecidable, whereas for ontologies formalised in DL it is generally decidable. I.e., given the ontologies O_1, O_2, and a signature morphism σ from O_1 to O_2, it is decidable whether the σ-translated axioms of O_1 are entailed by O_2. However, the combinatorial explosion yielded by trying to find all possible symbol mappings between two given ontologies makes such a brute force approach unpractical.

To obtain one of the answer sets above in reasonable time (i.e seconds or minutes), we necessarily have to relax our initial goal towards an approximation of the set of all possible contexts for a given ontology. In summary, our approach for

[1] This is a purely syntactic and somewhat lax definition which we adopt here only for lack of space; a more elaborate definition would be defined via a notion of logic translation using e.g. institution comorphims, see [Kutz *et al.*, 2008].

[2] E.g. the OneOf constructor found in many description logics allowing a finite enumeration of the elements of a concept is also expressible as a disjunction of nominals, and conversely. Such translations/simulations can be handled by a library of logic translations.

the first-order case[3] is based on formula matching modulo an equational theory—elaborated in detail in [Normann, 2009]. We want to outline this in the following.

Suppose we are given a source ontology O_1 and a target ontology O_2, which we assume have been translated to first-order via the standard translations. In the first step, we normalise each sentence of these ontologies according to a fixed equational theory. The underlying technique basically stems from term-rewriting: rewrite rules represent an equational theory such that all sentence transformations obtained through these rules are in fact equivalence transformations, e.g. such as $\neg A \sqcap \neg B \mapsto \neg(A \sqcup B)$. A normal form of a convergent rewrite system is then the unique representative of a whole equivalence class of sentences. The goal of normalisation is thus to identify (equivalent) expressions such as $\neg(\exists R.A \sqcap B)$ and $\neg B \sqcup \forall R.\neg A$.

In the next step, we try to translate each normalised axiom φ from O_1 into O_2, i.e. we seek a sentence ψ in O_2 and a translation σ such that $\sigma(\varphi) = \psi$. Note that potentially each axiom can be translated to several target sentences via different signature morphisms. To translate all axioms of O_1 into O_2, there must be a combination of *compatible* signature morphisms[4] determined from the previous, single sentence matchings. This task is also known as (consistent) many-to-many formula matching. In fact many-to-many formula matching modulo some equational theory is already applied in automated theorem proving (ATP) [Graf, 1996]. However, our approach is different in a crucial aspect: it allows for significant search speed up. We are normalising all ontologies as soon as they are inserted into the repository, i.e. not at cost of query time. Only the normalisation of the query ontology is at query time. Moreover, the normal forms not just allow for matching modulo some equational theory, but also enable a very efficient matching pre-filter based on skeleton comparison. A sentence skeleton is an expression where all (non-logical) symbols are replaced by placeholders. E.g., $\square \sqsubseteq \square \sqcup \square$ is the skeleton of $A \sqsubseteq B \sqcup C$. Obviously, two sentences can only match if they have an identical skeleton. Since syntactic identity can be checked in constant time, a skeleton comparison is a very efficient pre-filter for sentence matching.

All the presented techniques were developed in the context of formalised mathematics and a tool for the automated discovery of theory interpretations in first-order logic has already been implemented [Normann, 2009], and is currently being integrated into the Hets system [Mossakowski *et al.*, 2007]. This has been used for experiments on a **FOL** version of the Mizar library [MizarKB] that contains about 4.5 million formulae distributed in more than 45.000 theories, and thus is the world's largest corpus of formalised mathematics. Experiments where each theory was used as source theory for theory interpretation search in the rest of the library demonstrated the scalability of our approach. On average, a theory interpretation search takes about one second and yields 60 theory interpretations per source theory.

4 Discussion and Future Work

Because of the encouraging results in formalised mathematics, we are currently adopting and modifying these techniques for the application in the realm of ontologies. The techniques we have sketched above are directly applicable to two of the cases we have discussed: to searching for stringent contexts in the case where both ontologies are formalised in **FOL**, and to the case of conforming context where the logics can be in a DL or **FOL**, but where no definitional axioms are required, or where they are added manually. Automated search of such definitions is not yet supported.

Of course, there is no guarantee that what is successful for mathematical theories is equally successful for formal ontologies, and some of the characteristics and features regularly found in ontologies are problematic. For instance, ontologies are often formalised in DL as opposed to first- or higher-order logics used in formal mathematics. Hence, formal mathematical theories are in general constituted by much more complex axioms than formal ontologies (many ontologies have no other axioms than is-a hierarchies). The lower complexity of ontology axioms has the effect that many axioms share the same skeleton. This makes skeletons a less effective pre-filter, which means that the reduction of the search space for candidate signature morphisms will be less significant.

Initial experiments on DL ontologies already suggested some ideas on how to overcome these problems in future work:

- Interactive search space reduction: the user should be able to enforce some mappings of non-logical symbols—often some mappings are explicitly intended.
- Exploitation of the decidability of DLs.
- Specialised normal forms designed for various DLs.

Acknowledgements

We gratefully acknowledge the financial support of the European Commission through the OASIS project (Open Architecture for Accessible Services Integration and Standardisation) and the Deutsche Forschungsgemeinschaft through the Research Center on Spatial Cognition (SFB/TR 8). The authors would like to thank Joana Hois for fruitful discussions.

References

W. M. Farmer. Theory Interpretation in Simple Type Theory. In *Higher-Order Algebra, Logic, and Term Rewriting*, volume 816 of *LNCS*, pages 96–123. Springer, 1994.

P. Graf. *Term Indexing*, volume 1053 of *Lecture Notes in Computer Science*. Springer, 1996.

O. Kutz, D. Lücke, and T. Mossakowski. Heterogeneously Structured Ontologies—Integration, Connection, and Refinement. In *Advances in Ontologies (KROW-08)*, volume 90 of *CRPIT*, pages 41–50. ACS, 2008.

C. Meilicke, H. Stuckenschmidt, and A. Tamilin. Reasoning Support for Mapping Revision. *Journal of Logic and Computation*, 2008.

J. Meseguer. General logics. In *Logic Colloquium 87*, pages 275––329. North Holland, 1989.

Mizar Mathematical Library. `http://www.mizar.org/library`.

T. Mossakowski, C. Maeder, and K. Lüttich. The Heterogeneous Tool Set. In Orna Grumberg and Michael Huth, editors, *TACAS 2007*, volume 4424 of *LNCS*, pages 519–522. Springer, 2007.

I. Normann. *Automated Theory Interpretation*. PhD thesis, Department of Computer Science, Jacobs University, Bremen, 2009.

L. Serafini and P. Bouquet. Comparing Formal Theories of Context in AI. *Artificial Intelligence*, 155:41–67, 2004.

[3] In principle, these methods can be applied to any formalised content as long as the entailment relation obeys certain properties (as specified e.g. in entailment systems [Meseguer, 1989]).

[4] Two signature morphisms are compatible if they translate all their common symbols equally.