Using Ontologies to Model Polysemy in Lexical Resources

Fahad Khan[1] and Francesca Frontini[1, 2]

[1]Istituto di Linguistica Computazionale "A. Zampolli" , CNR Pisa
[2]Laboratoire d'informatique de Paris 6 , Labex OBVIL , Paris
firstname.secondname@ilc.cnr.it

Abstract

In this article we look at how the use of ontologies can assist in analysing polysemy in natural languages. We develop a model, the Lexical-Sense-Ontology model (LSO), to represent the interaction between a lexicon and ontology, based on *lemon*. We use the LSO model to show how default rules can be used to represent semi-productivity in polysemy as well as discussing the kinds of ontological information that are useful for studying polysemy.

1 Introduction

Given the current high levels of interest in lingustic linked open data and the availability of large scale, wide coverage ontologies like DBpedia and SUMO it was inevitable that there should also be an increased focus on the idea of using computational ontologies to provide semantic information for lexical resources, especially within the context of the Semantic Web. In this article we look at different ways in which ontologies and ontological knowledge can potentially help to describe and analyse the semantic phenomena of polysemy. Arguably the most popular RDF based model for linking together lexica with ontologies for the purpose of describing word meaning is *lemon* (Mccrae et al. (2012)). The *lemon* model is based on the principle of *semantics by reference* which foresees a clear separation of lexical and ontological layers in a lexico-semantic resource, using reified sense objects to map between the two, and argues for the semantics of a lexicon being wholly contained within the ontology (see Cimiano et al. (2013)). Our approach in this article is also based on a clear lexicon-ontology distinction in which senses are regarded as interfacing between lexical and ontological layers, as is posited in *lemon*. We will introduce our own model, the Lexicon-Sense-Ontology model (LSO) which is closely based on *lemon* (but which doesn't necessarily deal only with lexica and ontologies on the semantic web or only those resources represented in RDF) in Section 3. In Section 4 we use this model to investigate how best to exploit ontological information to represent cases of systematic polysemy while at the same time avoiding the problems raised by the sense enumeration lexicon model.

2 Ontology Modelling and Natural Language Meaning

If there were no real distinction to be made between ontological and semantic knowledge – or to be more accurate between how ontological and semantic knowledge is structured and arranged – it would be enough to link a lexical entry directly to an ontological vocabulary item. One could then use the inferential tools that have been developed for ontological representational languages like OWL to directly derive facts about, for example, synonymy and polysemy. It would then also be viable to treat a lexical resource like WordNet, which is essentially a semantic network hierarchically structured using lexical relations like hyponymy and meronymy, as an ontology (indeed WordNet has been used as an ontology in the past, although by now the limitations of such an approach have been made clear in works such as Gangemi et al. (2002) and Oltramari et al. (2002)). But there are in fact some important

differences between the two types of resource. Clarification on the differences in the arrangement and design of lexica and ontologies comes in the form of recent normative work on ontology design and especially via ontology evaluation methodologies. One of the most influential of these methodologies is OntoClean(Guarino and Welty (2004)). OntoClean provides a series of clearly formulated guidelines and suggestions for designing and evaulating ontologies based on well-studied metaphysical principles relating to such (technical) properties as *identity*, *rigidity*, *unity* – all of which turn out to be extremely salient for evaluating ontology design decisions. What is important for our purposes here is that the OntoClean principles are somewhat independent of those considerations based purely on language use and native speaker intuition that play a central role in the standard lexical semantic definitions of relations like hyponymy and synonymy. On the other hand the kinds of semantic information one would want to represent in a lexicon will include not just descriptions of the types of thing that a word refers to in the world, i.e., the extensional aspect of a word's meaning, but also information about language use, e.g., data about how and in what contexts a word can refer to different things. In addition, the association between an ontological term and its label is of a different kind from the association of a word sense or a word meaning with its head form[1]. For a more general overview on the differences between lexica and ontology see Hirst's survey article, Hirst (2004).

These observations about the relative language independence of well designed ontologies are important in justifying the use of ontologies as resources that enable researchers interested in natural language semantics to not only compare the meanings of terms across languages, using the ontology like an *interlingua*, but also to study how linguistic concepts map onto a (relatively) language independent domain. The fact that the meanings of ontological items tend to be comparatively "stable" and that ontologies are usually represented in formal languages for which there exist automated inference engines makes them extremely valuable in this respect. How then do we use ontological information to represent and to reason about semantic information? As we mentioned above, given the differences between semantic and ontological information, it's probably best not to treat ontological vocabulary items as word senses and link lexical entries directly to ontological items. At the same time, we want to access the information stored in the ontology in order to describe and reason about semantic information, although in a restricted way. Below we will describe a model for the interaction between a lexicon and an ontology in the modelling of semantic information, and show how it can be applied by focusing on the representation of the semantic phenomena of polysemy.

3 A Model of the Lexicon-Ontology interface

In this section we give a brief sketch of the model, the Lexicon-Sense-Ontology model (LSO), that we will use in the rest of the paper for representing the lexico-ontology interface. LSO is based on the *lemon* model, but with a number of alterations, especially in terms of how LSO represents word meaning as distributed across the lexicon and the ontology: for us a sense is not necessarily always to be regarded as a reified pairing of a lexical entry with an ontological entity, as in *lemon*. In the LSO model a lexicon Lex in a language $\mathfrak{L}$ is represented as a finite set $\{l_1, ..., l_k\}$ of lexical entries each of which can be tagged with morphosyntactic information and each of which is associated with one or more sense objects that represent the meaning of the lexical entry[2]. The sense relation $sense \subseteq Lex \times Sense$ relates lexical entries together with their senses. In LSO *homophonous* words like *bank* and *bank* exist as separate entries and we make the assumption that all of the senses linked to a single lexical entry by $sense$ are somehow related or have some kind of overlap between them. An ontology is a logical theory $\mathcal{O}$ in a logical language $\mathcal{L}$ with vocabulary $\mathcal{V}$. Members of the set $Sense$ are linked to ontological vocabulary items that describe the references of these senses using the relation $hasRef \subseteq Sense \times \mathcal{V}$. By overloading this relation we define $hasRef \subseteq Lex \times \mathcal{V}$ to represent the case where a given lexical

[1] C.f for example the discussion in Pease and Li (2010) on the difference between SUMO term names and lexical entries.

[2] We view senses as abstract representations of the meaning of a lexical entry, so that together a lexical entry and a corresponding sense form a kind of "form-meaning [complex] with (relatively) stable and discrete semantic properties which stand in meaning relations such as antonymy and hyponymy" Cruse (1986).

entry has a sense with a certain extension represented by a ontological vocabulary item (this is useful when we don't want to explicitly mention sense objects, as in the formulae in Section 4.2). In addition the relation $hasRefSub \subseteq Lex \times \mathcal{V}$ is used in case a given lexical entry has a sense with a certain reference (which may or may not be explicitly enumerated as a vocabulary item) that if it were to be a vocabulary item would be subsumed by a named ontology vocabulary item. As we noted above we do not regard senses as reified pairings between lexical entries and ontological entities because we leave open the possibility that a sense may not be mapped to a concept in an ontology – or at least not via the $hasRef$ relation. In contrast to *lemon* we do not consider the semantics of lexical entries to exist only in the ontology, as per semantics by reference, but that to a large extent (language specific) semantic data is also represented in the sense layer (which is for us part of the lexicon) and especially in the interrelationships between the sense objects and in the relationships between the lexicon and ontology. It was for these reasons that we decided not to re-use the already existing *lemon* model in this work, but to develop a slightly different one – particularly since *lemon* has a clear formal semantics which strongly limits the kinds of interpretations that one can make. In the LSO model we have essentially three layers, a morpho-syntactic layer, a sense layer, and an ontological layer[3]. The sense layer is language specific since different languages will map their senses onto the ontology in distinct and incompatible ways[4].

4 Representing Polysemy

One of the main advantages of the LSO model is that it can help to avoid some of the many pitfalls associated with what Pustejovsky calls the Sense Enumeration Lexicon (SEL) (see Pustejovsky (1995)). The term SEL is used to describe any lexicon in which the related meanings of a lexical entry are represented as a set of different senses each of which is stored separately and without any kind of additional structure to relate the senses together. The problem with this sort of arrangement is that it makes it difficult to account for the creativity in natural languages that allows language users to regularly use words in novel ways and still be understood. Such a simplistic model also renders it impractical to represent the various different shades of meaning that any single word or lexical entry may potentially have; with SELs we lose out on the relatedness between the different senses of the same lexical item. For example, the English word *school* can mean both a building as well as an institution. These are two different senses of the same word, and they are clearly closely related, but what is the best way to represent this relation? One plausible answer, as suggested by Generative Lexicon (GL) theory (Pustejovsky (1995)) is that different kinds of common sense or ontological information are more accessible to the entries in a lexicon than others and that they license different type of sense extension. One general strategy, then, for avoiding SELs is to allow lexical entries to systematically pick out certain aspects of ontological knowledge via sense objects in a way that allows the easy generation of additional meanings based on a limited and finite *stored* set of senses for each lexical entry. In the following sections we show how to model this kind of lexicon-ontology interaction and also how to represent polysemy using LSO.

4.1 Dealing with Semi-productivity in Polysemy

An important issuse to take into consideration here is that polysemy tends to be semi-productive and so an impoverished sense layer or even the lack of one would over generate instances of polysemy. For instance in Parole Simple Clips, a large scale wide coverage Italian lexicon (Lenci et al. (2000)), a polysemy alternation is recorded for proper nouns referring to locations between the types `HUMAN-GROUP` and `GEOPOLITICALLOCATION`: so that the name of a location like *Genova* can also name the inhabitants, or a representative group of inhabitants, from that location. However this rule doesn't apply to imaginary locations like `Eldorado` that in other linguistic respects seem to behave just like real locations. The

[3]With the first two of these layers comprising the lexicon.

[4]To put it crudely this tripartite division reflects a kind of rough division of labour between those linguists who concern themseleves with morpho-syntactic data; those linguists who concern themselves mostly with the peculiarities of natural language semantics; and finally ontology engineers.

PLANT-FRUIT alternation is well known and exists in many languages. In some languages, however, it interacts with a derivation rule. So that for instance in Italian many plants have masculine names whereas the fruit is feminine[5]. In order to know when the regular polysemy is acting without change of morphological gender, we need to allow for a reasonably complex interaction between lexicon and ontology that depends on factors such as whether the plant is relatively small in size or whether the fruit and tree are "exotic" to Italy. We can then say that this alternation is in fact limited to a fairly large subset of fruit plants that can be identified productively by accessing ontological knowledge.[6]

4.2 Using Default Rules

Polysemy alternations tend to be reasonably regular but admit of exceptions (which differ across languages) that can usually be enumerated as finite lists of exceptions or described using simple logical formulae that refer to ontological vocabulary items. This would suggest the use of a non-monotonic logic to represent polysemy in terms of ontological knowledge, and indeed, as we shall see below Reiter's Default Logic (Reiter (1987)) lends itself particularly well to this task. One should bear in mind, however, that regardless of the exact representation framework that we use or how these rules are implemented in an actual application, what is important here is to emphasise the use of information about natural language semantics to constrain *how* we access the ontological layer so that any application that uses the ontological data doesn't overgenerate 'examples' of polysemy.

Default logic is a popular non-monotonic knowledge representation language that uses rules to represent facts and statements that hold by default in addition to knowledge bases consisting of sets of first order logic or description logic formulae. Default rules are usually represented in the form $\frac{\phi:\psi_1,...,\psi_k}{\chi}$ – where the formula χ, the *consequent*, follows from ϕ, the *pre-requisite*, if it is consistent to assume $\psi_1, ..., \psi_k$, the *justifications*. In addition we can use classical rules to formulate the exceptions, that is, the cases when it's not acceptable to assume $\psi_1, ..., \psi_k$. A default theory is a pair consisting of a set of default rules, and a set of classical logic formulae. The semantics for default logic is usually given in terms of *extensions* which are sets of formulae with appropriate closure conditions that we won't describe here (Reiter (1987)). Default Logic is an appropriate formalism for cases where we are dealing with rules for which we do not know the set of exceptions beforehand or when it would be too difficult to enumerate or describe them all; the use of default logic also helps to emphasise that we are dealing with what is *usually* the case. So for example, take the ANIMAL-FOOD alternation, according to which the same word used to name an animal is also *usually* used to name the (edible) flesh of that animal. Say we are working with an English language lexicon and an ontology with the classes Animal and Edible, and the relation fleshOf, then given the lexical entry l, and ontology vocabulary items c, c', we can give the following default rule:

$$\frac{hasRef(l,c) \wedge c \sqsubseteq \texttt{Animal} \wedge \texttt{fleshOf}(c',c) \wedge c' \sqsubseteq \texttt{Edible} : hasRef(l,c')}{hasRef(l,c')}.$$

This rule is an example of a *normal default rule*, that is a rule where the justifications and the consequent are the same. We can read the rule above as saying that: if it is true that l can refer to the class c, a subclass of *Animal*, and if the flesh of the members of c, represented by the class c' is edible – then if it's consistent to assume that l has the extension c', we can indeed assume it to be the case. We can then add a (classical logic) rule such that lexical entries such as *Pig* and *Cow* do not name the (edible) flesh of the animals referred to by those nouns in English. So that if $l = Pig$, then it is not consistent to assume that l can also mean the flesh of a pig. In effect then, through the implementation of such default rules, we can use the ontological layer of a lexico-semantic resource modelled using the LSO model to justify polysemy alternations. In the example we gave above two things have the same name because one is a part of the other – and we can check using the ontology what this part_of relation actually consists in. This is why it's important to be able to make a clear distinction between what is in the ontology and

[5]For example, apple tree and apple are *melo* and *mela* respectively.

[6]See Copestake and Briscoe (1995) for an interesting discussion of the contextual blocking effects of both lexical and ontological knowledge on polysemy rules.

what is in the lexicon in order to avoid the danger of circularity in these explanations. The "messy" semantic details of how a language like English represents the relationship between animals and their flesh, the socio-cultural reasons as to why *Beef* or *Pork* are used instead of *Cow* and *Pig*, and that serve to somehow "distort" the ontological data, are part of the structure of the sense layer. It is especially important to emphasise this since there are languages such as West Greenlandic Eskimo in which the kind of "grinding" phenomena discussed above doesn't occur (Nunberg and Zaenen (1992)). The benefit of having a relation like $hasRefSub$ is that we don't need to explicitly store senses and this can be very useful. For example, according to OntoClean principles the class of Animals is not subsumed by the class of Physical Objects instead there exists a different (part_of) relation linking an animal with its physical body. But the large majority of languages do not seem to lexicalise this difference, and so the following rule can be justified for most lexica: given $l \in Lex, c \in \mathcal{V}$, then

$$hasRef(l, c) \wedge c \sqsubseteq \texttt{Animal} \rightarrow hasRefSub(l, \texttt{PhysicalObject}).$$

This ability to refer to senses without storing them at least partially obviates some of the problems inherent in SELs. It also means that we can distinguish cases when a sense really does have an existing ontology class as an extension (this is especially true in technical and scientific contexts or in controlled versions of natural languages); on the other hand it may be that the ontology we're using doesn't contain a specific concept, e.g., there may not be an ontology item corresponding to the fruit Persimmon, necessitating that the word sense in question be linked to the more general class `Fruit` using $hasRefSub$.

With this kind of semantic-ontological information available we can easily construct systems for word sense disambiguation that can capture cases of polysemy by keeping track of the kinds of ontological knowledge that lead to polysemy while at the same time avoiding overgeneration by storing exceptions to the rules. The problem is how to implement the default rules themselves. The idea of extending description logics with default rules hit a major stumble due to the undecidability result in Baader and Hollunder (1995) – although they did show that decidability was preserved in the case of formulae with named individuals. In certain limited cases, however, such as for example the extension of description logics with normal default rules using a special kind of default semantics, decidability is preserved, but further work needs to be done in order to study the extent to which this will enable us to capture the kinds of semantic information that we want to represent (see Sengupta et al. (2014)). There are also several other ways of integrating description logic databases with default rules: see for example the work of Dao-Tran et al. (2009) which makes use of conjunctive query programs. Further work in this area will look into the best combination of formalism and efficient knowledge representation tools in order to represent natural language semantics using the LSO model.

4.3 Further Observations on Polysemy and the Structuring of the Sense Layer

One issue that commonly arises when trying to model polysemy phenomena using ontologies concerns the need to have access to knowledge about what is *usually* the case in both the physical world and in social reality; and here one should stress the importance of ontologies that deal with social reality with respect to this task. Polysemy occurs in contexts where the association between two or more entities or aspects of the same entity is strong enough that the advantage gained by using the same term to refer to both of them outweighs whatever risk there may be of confusion[7]. In this section we look at how such ontological knowledge can be useful in interpreting polysemy. For instance one particularly interesting class of examples of polysemy is that relating to lexical entries that name both information objects and physical objects such as *book*, *play*, *poem*, and *film*. For instance take the lexical entry *book* that can mean both `BOOK_AS_PHYSICALOBJECT` and `BOOK_AS_INFORMATIONOBJECT` as in the sentences

- *The book had yellowed with age and gave off a musty odour* and
- *The book was thrilling and suspenseful.*

[7]Obviously the trade-off varies with context, so that referring to a customer as "the ham sandwich" is a viable communication strategy in a deli.

It is unlikely that the concept BOOK_AS_PHYSICALOBJECT will be explicitly modelled in most ontologies and so *hasRefSub* comes in useful again. The sense of *book* in which it refers to an information object seems to be primary here; books are informational objects that are *usually* instantiated as physical objects or are *usually* stored in some kind of physical format[8]. On the other hand lectures are not, or at least not by default, published but are instead more closely associated with events and informational objects.

- *?The lecture is lying on the table.*
- *The lecture is on my hard drive. / The lecture took an hour. / The lecture was both enthralling and informative.*

The first sentence in the preceding sounds slightly odd, but is still understandable since lectures are often instantiated as sets of notes or occasionally published as books (instead, a sentence like *The lecture notes are lying on the table* is much more acceptable); the second instance is much more acceptable because of the common practice of storing footage of lectures or the slides used in a lecture in a digital format; the final two are both completely acceptable. Other informational objects like conversations and speeches are not associated with any particular physical format by default, and can only in special contexts be considered as acceptable arguments with predicates that select for physical object arguments, although they are much more acceptable with predicates that select for digital or analogue data objects. Another important issue when dealing with polysemy is to determine when one sense of a polysemic word is somehow more primary or *established* to use Cruse's terminology in Cruse (1986); this in turn would suggest some further structuring in the sense layer to account for the distributions of senses. To take the example given in Cruse (1986):

- *I'm not interested in the cover design, or the binding – I'm interested in the novel.*
- *?I'm not interested in the plot, or the characterisation, or anything of that nature – I'm interested in the novel.*

This example is also productive in that it holds for films when stored in DVDs (e.g., *I'm not interested in the case design or the booklet – I'm interested in the movie.*?*I'm not interested in the plot, or the acting or anything like that – I'm interested in the movie.*).The established or primary status of certain senses of a word is useful information that can go in the sense layer since it does affect how a word behaves.

5 Conclusions and Further Work

Lexical semanticists have studied the lexicon-ontology interface for many years, investigating the best way to divide up semantic and ontological information at the level of theory. Nowadays, thanks in large part to the popularity of the linked data movement, we actually have the possibility of accessing large-scale wide-coverage ontologies that are comprehensive enough to study these more general theories of the lexicon using computers; at the same time ontology engineering is maturing as a discipline and has also been able to contribute a great deal to the debate in its own turn. In this article we have attempted to introduce a general framework to study some of these issues. What's clear however is that most existing ontologies are not designed according to the strict constraints described in the OntoClean model – and that many of them do in fact make the kinds of confusions between lexical and ontological information that we alluded to above. However we still feel that enough of a distinction is observed in practice to render our work useful in the context of lexicon-ontology interfacing, even if it's as an idealisation. We are currently in the process both of enriching our model in order to describe diverse types of semantic information and of determining how to actually implement some of the ideas introduced in this paper using currently available lexicons and ontologies. In future we plan to place a greater emphasis in our research on producing resources and tools, e.g., lexical databases that answer given queries either by searching among existing senses or by triggering the correct rules to produce senses on the fly.

[8]Once more Default Logic again seems to be the obvious choice to represent these kinds of facts.

References

Baader, F. and B. Hollunder (1995). Embedding defaults into terminological representation systems. *J. Automated Reasoning 14*, 149–180.

Cimiano, P., J. McCrae, P. Buitelaar, and E. Montiel-Ponsoda (2013). On the role of senses in the ontology-lexicon. In *New Trends of Research in Ontologies and Lexical Resources*, pp. 43–62. Springer.

Copestake, A. and T. Briscoe (1995). Semi-productive polysemy and sense extension. *Journal of semantics 12*(1), 15–67.

Cruse, D. (1986). *Lexical Semantics*. Cambridge, UK: Cambridge University Press.

Dao-Tran, M., T. Eiter, and T. Krennwallner (2009). Realizing default logic over description logic knowledge bases. In C. Sossai and G. Chemello (Eds.), *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 10th European Conference, ECSQARU 2009, Verona, Italy, July 1-3, 2009. Proceedings*, Volume 5590 of *Lecture Notes in Computer Science*, pp. 602–613. Springer.

Gangemi, A., N. Guarino, A. Oltramari, R. Oltramari, and S. Borgo (2002). Cleaning-up wordnet's top-level. In *In Proc. of the 1st International WordNet Conference*.

Guarino, N. and C. A. Welty (2004). An overview of ontoclean. See Staab and Studer (2004), pp. 151–172.

Hirst, G. (2004). Ontology and the lexicon. See Staab and Studer (2004), pp. 209–230.

Lenci, A., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, and A. Zampolli (2000). SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography 13*(4), 249–263.

Mccrae, J., G. Aguado-De-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner (2012, December). Interchanging lexical resources on the semantic web. *Lang. Resour. Eval. 46*(4), 701–719.

Nunberg, G. and A. Zaenen (1992). Systematic polysemy in lexicology and lexicography. In *EURALEX'92. Papers submitted to the 5th EURALEX International Congress of Lexicography*.

Oltramari, A., A. Gangemi, N. Guarino, and C. Masolo (2002). Restructuring WordNet's Top-Level: The OntoClean approach. In *Proceedings of LREC2002 (OntoLex workshop). Las Palmas, Spain*.

Pease, A. and J. Li (2010). Controlled English to Logic Translation. In R. Poli, M. Healy, and A. Kameas (Eds.), *Theory and Applications of Ontology*. Springer.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Reiter, R. (1987). Readings in nonmonotonic reasoning. Chapter A Logic for Default Reasoning, pp. 68–93. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Sengupta, K., P. Hitzler, and K. Janowicz (2014). Revisiting default description logics and their role in aligning ontologies.

Staab, S. and R. Studer (Eds.) (2004). *Handbook on Ontologies*. International Handbooks on Information Systems. Springer.