Polysemy in Controlled Natural Language Texts

Normunds Grūzītis and Guntis Bārzdiņš

Institute of Mathematics and Computer Science, University of Latvia
normundsg@ailab.lv, guntis@latnet.lv

Abstract. Controlled natural languages (CNL) and computational semantics in general do not address word sense disambiguation, i.e., they tend to interpret only some functional words that are crucial in the construction process of discourse representation structures. We present two alternative frameworks for supporting polysemy in controlled languages. The approaches result in more natural CNLs suitable for description and translation of multi-domain texts.

1 Introduction

There are several sophisticated controlled natural languages (CNL), which cover relatively large subsets of English grammar providing seemingly informal means for knowledge representation (Schwitter et.al., 2008). CNLs typically support bidirectional mapping to some formal language like first-order logic (FOL) or its decidable subset OWL DL (Kaljurand and Fuchs, 2006) that allows to apply existing tools for reasoning, consistency checking or even basic satisfiability model building.

Two commonly accepted restrictions are used in CNLs to enable construction of unambiguous discourse representation structures (DRS): a set of interpretation rules for potentially ambiguous syntactic constructions, and a monosemous lexicon — words are treated as predicate identifiers whose 'meaning' is defined only by FOL formulas derived from the text being analyzed. While the first restriction limits only syntactic sophistication of a language, the second one causes essential communication limitations as the natural language lexicon is inherently polysemous (e.g. "The library$_{\text{[collection]}}$ consists of million books"; "The city is constructing a new library$_{\text{[building]}}$"). The polysemy of natural language is not a deficiency, but rather a gateway for referring to the rich background knowledge invoked by the same lexemes in different contexts thus leading to multiple word-senses. In this paper we address the latter limitation.

The root cause of polysemy in a natural language is that at any given moment there is only a 'finite' number of lexemes, which speakers of the given language know and thus can use for communication. Meanwhile there is a potentially unlimited number of new concepts (discourses) that might need to be named and communicated about. Metaphoric reuse of existing lexemes therefore is unavoidable in the natural language, which can be summed in a saying: language is a graveyard of 'dead' metaphors (Leary, 1994). Fortunately, various metaphoric senses of the same lexeme typically fall in radically different domains, which is helpful in word sense disambiguation (WSD).

2 OWL DL compliant micro-ontologies as interlingua

A monosemous lexicon (terminology), of course, is appropriate for descriptions that verbalize single-domain knowledge (i.e., consistent OWL DL ontologies). However, even seemingly consistent descriptions might need to be artificially split into two or more micro-domain descriptions to avoid lexical ambiguities and to maintain compliance with existing naming conventions. A possible alternative in such cases is to introduce artificial lexemes by explicitly pointing out the specific meanings (e.g. “library-building” versus “library-collection”), but then the language becomes rather un-natural and dependent on specific domain-ontology naming.

Internally consistent domain ontologies that follow lexicon-driven naming conventions we call *micro-ontologies*. Table 1 illustrates WSD problem when the text references different domains that use overlapping and potentially inconsistent terminology.

Table 1. Sample snippets from three domain ontologies verbalized in Attempto Controlled English. If in a text the polysemous lexeme “library” is used to reference building and collection domains alternately, then the appropriate sense has to be implicitly assigned to each utterance in order to consistently merge the assertion with the appropriate domain ontology.

	Micro-ontologies (ontological text)	
T-Box	**Domain**	**Axioms in ACE**
	Buildings	*Every building is a construction and has a roof.* *Every library is a building.*
	Collections	*Every collection is an abstract-entity that contains some items.* *Every library is a collection that contains some publications.*
	General	*Every construction is a physical-entity.* *No physical-entity is an abstract-entity.*
A-Box	**Assertions** (factual text)	
	There is a library$_{\text{[buildings]}}$ *that has a green roof.* *The library*$_{\text{[collection]}}$ *contains some valuable publications.*	

The role of polysemy is most clearly apparent from the multilingual point of a view: it is impossible to avoid interpretation of lexemes when translating a text. In our case, *interpretation* means selection of the appropriate micro-ontology (see Table 2).

Although grammar constructions (OWL DL mappings) and lexicons for the source and target languages would differ, the *interlingua* — OWL DL micro-ontologies and their consistent *mergers* — remain the same. Moreover, by attaching translation equivalents to the ontological concepts, micro-ontologies simultaneously serve as monosemous multi-lingual lexicons facilitating the translation process. The term *interlingua* we mean in a wider sense: not only in the multi-lingual context, but also for monolingual multi-domain communication.

The problems of WSD and ontology merging are tightly intertwined and, in our view, the lack of definitive success is largely due to addressing these issues separately. Therefore we address both of these problems simultaneously — OWL DL formal semantics can be used to dynamically handle micro-ontologies for WSD over polysemous factual sentences. From all the available micro-ontologies for each sentence (or clause) are selected those that can be invoked (directly or via some merger) by a target lexeme (typically, a predicate) or other syntactically related lexemes (syntactic links are mapped onto ontology properties). In general, more than one micro-ontology can be invoked by an assertion due to the fact that different word senses are not necessarily mutually inconsistent. Selecting the largest micro-ontology merger likely unveils the most specific meaning (and facilitates further reasoning tasks). However, it is not necessary to get rid of the consistent alternatives — in case of later inconsistency they can be used during backtracking. As long as the current discourse remains consistent, its merged ontology is gradually augmented; otherwise additional discourse ontology is created separately.

Table 2. Polysemy in multilingual communication. The two example sentences can be correctly translated (interpreted) by consistently merging the appropriate domain ontologies.

	Micro-ontologies (ontological text)	
	Domain	**Axioms in FOL**
T-Box	#1	$\forall x(\text{artifact}(x) \rightarrow \neg\text{body-part}(x))$ $\forall x(\text{footwear}(x) \rightarrow \text{artifact}(x))$
	#2	$\forall x(\text{shoe}(x) \rightarrow \text{footwear}(x))$ $\forall xy(\text{polish}(x,y) \rightarrow \text{person}(x)\ \&\ \text{footwear}(y))$
	#3	$\forall x(\text{nail}(x) \rightarrow \text{body-part}(x))$ $\forall xy(\text{polish}(x,y) \rightarrow \text{person}(x)\ \&\ \text{nail}(y))$
A-Box	**Assertions** (factual text)	
	Source text (EN)	**Target text** (LV)
	John polishes$_{[2]}$ *a shoe.* *Ann polishes*$_{[3]}$ *some red nails.*	*Jānis pucē*$_{[1 \oplus 2]}$ *vienu kurpi.* *Anna vīlē*$_{[1 \oplus 3]}$ *sarkanus nagus.*

The proposed concept of micro-ontology is similar to the Cyc concept of micro-theories (Lenat, 1995), where all world-knowledge is split into narrow domain micro-theories (ontologies). In our approach micro-ontology is one of many internally consistent domain-ontologies (or their dynamic mergers) described in OWL DL (through a CNL or directly in an ontology editor), against which the polysemous lexemes of the factual sentences can be mapped.

3 Alternative approach to polysemy and discourse in CNL

The above proposed rather 'classic' solution for adding polysemy to CNLs is theoretically plausible, but it also raises a critical question: is this really how the natural language works? It is well acknowledged in linguistic and cognitive sciences that polysemes are etymologically and therefore semantically related, and typically originate from metaphoric usage (Ravin and Leacock, 2000). The metaphoric view erases the strict borders between polysemous word senses — these borders are shifting with each creative use of a metaphor, and dictionaries or ontologies shall be viewed only as short-lived snapshots of currently common word usages. To illustrate, a metaphoric statement "She is a star" in natural language implies only that a person possesses some aspect (e.g. being prominent) of a true star. Meanwhile a monosemous CNL would likely interpret it literally as a light-emitting celestial star.

Frame-semantic linguistic theory (Fillmore et.al., 2003) has already come up with a way to avoid such 'tyranny' of literal word meanings. Through extensive corpus analysis FrameNet has identified approximately 700 *frames* which can be invoked by actual words or sentences — regardless of being used literally or metaphorically. Translation of the input text into FrameNet frames would resolve the problem of polysemy. A CNL could help with translation disambiguation as explained below.

The ultimate purpose of a CNL is to build a formal DRS capturing the full semantics of the input text. Although one could try to merge the classic DRS construction techniques rooted in FOL with FrameNet for a more natural polysemous CNL, this would not aid the disambiguation problem. Therefore we propose an alternative discourse representation approach based on PDDL (Planning Domain Description Language) leading to a new kind of CNLs not rooted in FOL anymore.

PDDL (McDermott et.al., 1998) is designed to formalize dynamic models, where *actions* guide the model through a series of *states* — in contrast to static models specified by FOL. But most importantly — PDDL maps directly to FrameNet: a PDDL *action* in most cases is the same FrameNet *frame*. Thus PDDL adds a formal structure FrameNet was lacking — it introduces strict object and event identification and therefore allows for syntactic subordination and global co-referential anaphora encoding that is crucial for building large discourse structures. PDDL also comes with a powerful constraint mechanism — actions in PDDL have formal *precondition* and *effect*, which must be coordinated in consecutive actions to achieve a valid discourse model. These PDDL action constraints along with global anaphora resolution enable disambiguation of the FrameNet frame to be invoked by the particular lexeme.

Table 3. Alternative CNL discourse construction stages.

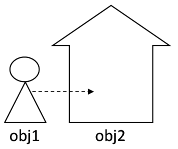
Text	FrameNet frame	PDDL	Text-to-scene
she	People	obj1 (anaphoric ref. to known object)	
enters	Arriving	:action ARRIVING :parameters (?theme ?goal) :precondition (not (in ?theme ?goal)) :effect (in ?theme ?goal)	
studio	Building_subparts	obj2	

Table 3 illustrates the main steps of the proposed approach. An interesting early observation: use of PDDL enables rather straightforward text-to-scene conversion — a discourse representation approach recently studied also by Johansson et.al. (2005).

4 Discussion

The idea to differentiate two kinds of sentences in natural language — the ontological and the factual ones (T-Box and A-Box in Tables 1 and 2) — turns out to be a helpful principle. Although natural language expressions occasionally might be a mixture of both kinds of sentences, mostly such distinction on sentence level is possible.

Polysemy is less relevant for the background knowledge (ontological sentences), which essentially define language-independent abstract concepts in some, usually monosemous, domain ontology. Meanwhile polysemy support is crucial for the factual communication, which typically does not explicitly reference the background knowledge (T-Box), which needs to be guess-mapped through the polysemous words used in the text. In a CNL the corresponding T-Box has to be introduced explicitly along with an A-Box, usually by manual sharing of ontological sentences among factual texts. This forbids possibility for (inconsistent) polysemy in existing CNLs.

While it is disputable whether a CNL is a more convenient approach for describing ontologies (T-Boxes) than the formal languages and their graphic editors, a CNL is definitely an advantage when describing concrete situations through factual sentences. Vast amounts of such descriptions already exist in a written form: consider, for example, information extraction from a newspaper archive, which is predominantly a factual text.

In contrast to the universal macro-ontologies, micro-ontologies offer several significant advantages: (a) they do not impose a single consistent scheme, allowing many distinct points of view to co-exist; (b) they can be seen as snapshots of some aspects of 'reality', supporting non-stable and temporal entities — existing ontologies don't have to be updated each time the reality changes; alternative ontologies should be introduced instead — it is a task of a reasoner to choose the appropriate ones; and (c) they scale well — things don't have to be compressed in a restricted number of categories thus avoiding 'signal losses'; the only restriction is the size of a lexicon.

In Sections 2 and 3 we have proposed two very different approaches for adding polysemy to CNLs — the micro-ontology approach in Section 2 is more traditional and compatible with existing CNLs, while the PDDL approach in Section 3 is more radical.

References

Fillmore C. J., Johnson C. R. and Petruck M. R. L. *Background to FrameNet*. In International Journal of Lexicography 16, 2003

Johansson R., Berglund A., Danielsson M. and Nugues P. *Automatic text-to-scene conversion in the traffic accident domain*. In Proceedings of the 19th International Joint Conference on Artificial Intelligence, 2005

Kaljurand K. and Fuchs N. E. *Bidirectional mapping between OWL DL and Attempto Controlled English*. In Proceedings of the 4th Workshop on Principles and Practice of Semantic Web Reasoning, 2006

Leary D. E. (Ed.) *Metaphors in the history of psychology*. Cambridge University Press, 1994

Lenat D. *Cyc: A Large-Scale Investment in Knowledge Infrastructure*. In Communications of the ACM, 38:11, 1995

McDermott D. et al. *PDDL — The Planning Domain Definition Language*. Technical report. Yale University, 1998. Available at: http://www.cs.yale.edu/homes/dvm/ [13/03/2009]

Ravin Y. and Leacock C. *Polysemy*. Oxford University Press, 2000

Schwitter R., Kaljurand K., Cregan A., Dolbear C. and Hart G. *A Comparison of three Controlled Natural Languages for OWL 1.1*. In Proceedings of the 4th International Workshop on OWL Experiences and Directions, 2008