Towards Structural Natural Language Formalization: Mapping Discourse to Controlled Natural Language

Nicholas H. Kirk
Computer Science Department
Technische Universität München
nicholas.kirk@tum.de

Abstract

The author describes a conceptual study towards mapping grounded natural language discourse representation structures to instances of controlled language statements. This can be achieved via a pipeline of preexisting state of the art technologies, namely natural language syntax to semantic discourse mapping, and a reduction of the latter to controlled language discourse, given a set of previously learnt reduction rules. Concludingly a description on evaluation, potential and limitations for ontology-based reasoning is presented.

1 Motivation

Work towards the formalization of natural language has been pursued on both syntactic and semantic levels. Controlled Natural Languages (CNL) for instance provide an unambiguous set of syntactic rules and a controlled vocabulary (Wyner et al., 2010), while sharing human intelligibility with the original Natural Language (NL) from which it derives (Kuhn, 2013). Approaches to pure semantic formalization have been done via symbolic and distributional characterizations (Blackburn et al., 2001; Harris, 1981), to various extents of compositionality (Clarke, 2012).

An important and structural approach towards formalization of discourse is Discourse Representation Theory (DRT) (Kamp, 1981; Kamp and Reyle, 1993), which makes use of inter- and intra-sentence discourse referents for anaphoric referencing and meaning preservation, and a set of semantic-level constraints over them. DRT maintains transformations to and from logic formalisms (Kamp and Reyle, 1993), and has direct applications within the automated sentence construction domain (Guenthner and Lehmann, 1984; Fuchs et al., 2010). Given the logical and linguistic properties of CNL (e.g. reasoning, paraphrasability, human- and machine- readability) the author stresses that a successful mapping between NL and CNL can enable language based cognition of simple autonomous software assistants, for reasoning and as interface to both peers and humans.

2 Concept

Given such rationale, the community should formulate a methodology for operating a reduction of sentence-level natural language discourse, to a discourse representation formulated in a target controlled natural language.

The author presents a possible pipeline abstraction of preexisting state-of-the-art means, as described in Figure 1. In particular, source channel text normalization (C1) to regularize erroneous phonetic transcriptions and spelling; a text to grounded Discourse Representation Structures (DRS) parser (C2) which works thanks to Combinatory Categorial Grammar (CCG), i.e. a grammar formalism that allows a computationally efficient interface between syntax and structural semantics (Curran et al., 2007). The implemented form has already achieved optimal results and can produce Discourse Representation Structures as output (Bos, 2008); a previously trained sentence-level Support Vector Machine (SVM) rule classifier, which identifies the types of NL to CNL reductions that should be operated (C3). A similarly implemented classifier is present in literature (Naughton et al., 2010). We then have a syntactic manipulation engine to transform the natural language input DRS into a set of compliant CNL DRS instances (C4), subject to the previously obtained classification results. Such classification (C3) should account for, for instance:

- intrinsically ambiguous natural language

Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue, December 16-18, 2013, Amsterdam, The Netherlands.

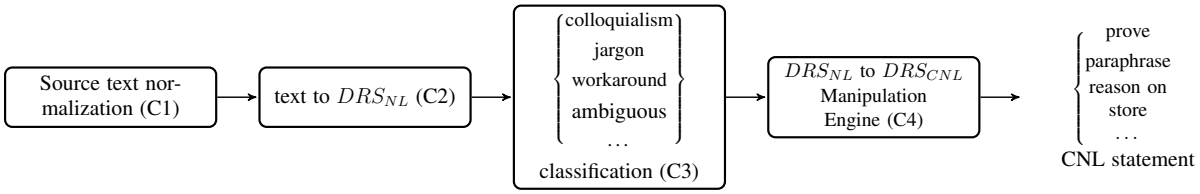

Figure 1: Representation of an abstract structure-level only NL to CNL manipulator

syntactic constructs

- ambiguous anaphoric reference resolution
- conscious constraining decisions on the expressiveness of specific CNL constructs

The full enumeration of reduction case reasons is application domain-dependent and require an apriori-istic study that can be performed online and in a supervised manner, for instance with active learning techniques. A possible target CNL which has proven robustness and reliability is ACE (Fuchs et al., 2006), which has DRS to CNL verbalization functionalities, as well as paraphrasing, proving and inference reasoning capabilities. Figure 2 shows a simple instance of the presented pipeline, which requires manipulation via sostitution of the unigram "linguistics" with the trigram "a linguistic class".

NL: "Harris can teach linguistics on Tuesdays."

⇓

ACE: "Harris can teach a linguistic class on Tuesday."

Figure 2: Example of an NL sentence instance and a possible semantic-preserving reduction to ACE

Evaluation Evaluation should mainly assess, via the use of human evaluation, if given an arbitrary sentence related to the application domain, the meaning of this has been successfully conveyed to the target controlled sentence. For instance, a threshold of satisfactory quality in action-oriented tasking domains (Nyga and Beetz, 2012) can be if arguments of intra-, mono-, ditransitive verb arguments have been preserved, together with correct anaphoric resolution. Evaluation will also assess domain-specific classification rates and computational efficiency.

Limitations The presented architecture does not make assumptions on the content of the predicates that are represented by words, given that the manipulation is operated only at a structural level, i.e. within the boundaries of DRS expressiveness. For a deeper predicate-related alignment, further considerations regarding lexicon should be made, to provide word sense and Part-Of-Speech (POS) mappings between source vocabulary and target controlled vocabulary.

Potential Current statistic-based web search approaches that make use of word n-gram models can exploit a more structural, discourse oriented approach. Formalization enables logic satisfiability check of manipulated NL questions via reduction and reasoning on First Order Logic (FOL) clauses. The expressiveness of the latter would also allow reasoning as Constraint Satisfaction Problems (CSP), i.e. a widely adopted mathematical formalism that expresses real-world decision problems as unary and binary constraints over finite variable domains. To pursue the example in Figure 2, admitting other ontological knowledge of lecturers' availability and ability, we could formulate an NL question (that becomes a formal ACE question) to ask for solutions to a simple timetable scheduling CSP problem, where the domains are the possible lecture days and types, and the constraints are the required lecture types and time precedence relations between them.

3 Future Work and Conclusions

This concept-only presentation hopes to have briefly highlighted the potential that such abstract CNL-based architecture can have, above all within the context of artificial assistants, as a means of interface, logic and combinatorial problem reasoning in ontology-based applications. If compliant with CNL rules, a specific set of syntactically reduced NL statements can seamlessly interface humans and machines while maintaining intelligibility and logical properties, such as entailment verification and inference. Future work should focus on implementation and efficiency verification of the stated architecture, to then investigate predicate-level (lexical) semantic align-

ment, to step towards (quasi-) complete sentence-level natural language formalization.

References

Patrick Blackburn, Johan Bos, Michael Kohlhase, and Hans De Nivelle. 2001. Inference and computational semantics. In *Computing Meaning*, pages 11–28. Springer.

Johan Bos. 2008. Wide-coverage semantic analysis with boxer. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, pages 277–286. College Publications.

Daoud Clarke. 2012. A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, 38(1):41–71.

James R Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 33–36. Association for Computational Linguistics.

Norbert E. Fuchs, Kaarel Kaljurand, and Gerold Schneider. 2006. Attempto Controlled English Meets the Challenges of Knowledge Representation, Reasoning, Interoperability and User Interfaces. In *FLAIRS 2006*.

Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. 2010. Discourse Representation Structures for ACE 6.6. Technical Report ifi-2010.0010, Department of Informatics, University of Zurich, Zurich, Switzerland.

Franz Guenthner and Hubert Lehmann. 1984. Automatic construction of discourse representation structures. In *Proceedings of the 10th international conference on Computational linguistics*, pages 398–401. Association for Computational Linguistics.

Zellig S Harris. 1981. *Distributional structure*. Springer.

Hans Kamp and Uwe Reyle. 1993. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Number 42. Springer.

Hans Kamp. 1981. A theory of truth and semantic representation. *Formal semantics-the essential readings*, pages 189–222.

Tobias Kuhn. 2013. The understandability of owl statements in controlled english. *Semantic Web*, 4(1):101–115.

Martina Naughton, Nicola Stokes, and Joe Carthy. 2010. Sentence-level event classification in unstructured texts. *Information retrieval*, 13(2):132–156.

Daniel Nyga and Michael Beetz. 2012. Everything robots always wanted to know about housework (but were afraid to ask). In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vilamoura, Portugal, October, 7–12.

Adam Wyner, Krasimir Angelov, Guntis Barzdins, Danica Damljanovic, Brian Davis, Norbert Fuchs, Stefan Hoefler, Ken Jones, Kaarel Kaljurand, Tobias Kuhn, et al. 2010. On controlled natural languages: Properties and prospects. In *Controlled Natural Language*, pages 281–289. Springer.