Natural Language Directed Inference in the Presentation of Ontologies

Chris Mellish, Department of Computing Science

The standard tool for articulating the shared assumptions of a community for knowledge representation in a domain is the *ontology*. An ontology is a list of standard terms to be used (usually based on words or phrases in some natural language), together with logical axioms about their intended interpretation (e.g. two terms might name classes, one of which is a subclass of the other or which are disjoint). Ordinary users of e-services will need to be able to understand the ontologies they are using whilst being insulated from the details of the logic sitting in the background. This project seeks to develop a generic approach for building *natural language* presentations of parts of ontologies. The project is supported by EPSRC grant GR/S62932.

Although ontology terms need not have any resemblance to natural language words, an ontology is not unlike a controlled natural language terminology. Therefore it is revealing to investigate the sorts of questions people ask when attempting to use such terminologies. We collected .ABs from web sites concerned with technical vocabulary and terminology. We found 13 appropriate web sites, yielding a total of 122 questions. Of these, 67% were simple *What is an X?* type questions. A further 11% asked about several related concepts in one question. The only other significant pattern (7%) was questions asking for a comparison. This initial investigation confirms our intention to start with a focus on *What is X?* type questions.

Natural Language Directed Inference

Answering *What is X?* involves presenting ontology axioms relevant to X. =owever there can be a "generation gap" (>eteer 1992) between ontology axioms and single natural language sentences - an axiom does not always package information appropriately for a single sentence. It could allow false implicatures when realised and it might result in an inelegant shifting of focus in the text.

To overcome these limitations, content determination must be able to select material in more ways than just choosing an axiom. Information expressed must be true, and so content determination will be a form of *inference* from the axioms. We call this *natural language directed inference* (NLDI). It is a kind of forwards inference which must satisfy eight main requirements. These requirements can be regarded as cases of the Gricean maxims for cooperative conversation (Grice 1975).

Grice, H. P. "Logic and Conversation". In P. Cole and J. Morgan (eds) *Syntax and Semantics* Vol 3, Academic Press, 1975.

Meteer, M., *Expressibility and the Problem of Efficient Text Planning*, Pinter, 1992.

Eight Requirements for NLDI

Inferred propositions must:

- Soundness: follow from the original logical theory (set of axioms)
- Relevance: contribute information relevant to the question being answered.
- Conservatism: be not very different from the original axioms
- Complexity: have appropriate linguistic complexity
- Coherence: satisfy linguistic coherence constraints (i.e. be linked to other selected material)
- Novelty: not have already been expressed (and not be tautologies).
- Fullness: be complete, to the extent that they don't support false implicatures
- User-orientation: be in accord with user model preferences

Techniques for NLDI

Refutation-based approaches to inference rely on having a precisely specified inference goal, whose negation is incompatible with the axioms. .or DLs, the standard tableaux method have similar properties. NLDI does not have an inference goal that can be expressed in structural terms. It is more akin to "non-standard" types of inference, perhaps to approximation (Brandt et al 2002), though again the target logical language is without a simple formal characterisation. Perhaps the closest approach we are aware of is meta-level control of inference, where factors outside of the logic (e.g. other kinds of descriptions of the shapes of logical formulae) are used to guide inference (Bundy and Welham 19A1).

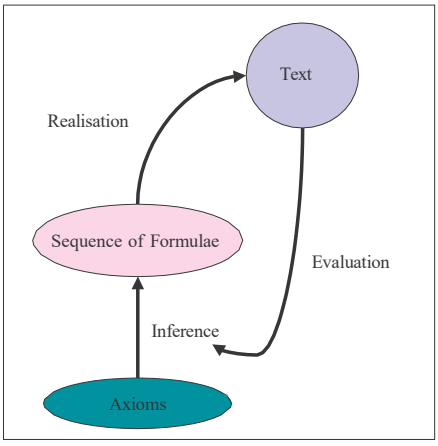
Brandt, S., KAsters, R. and Turhan, A.-+., "Approximation and difference in description logics". In *Procs of KR-02*, 2002.

Bundy, A. and Welham, B., "Using meta-level inference for selective application of multiple rewrite rule sets in algebraic manipulation", Artificial Intelligence 1B(2), 19C1.

An "overgeneration" architecture

The system we are developing has an "overgeneration" architecture, where multiple possible logical forms are proposed and then evaluated according to the linguistic properties of their best realisations. This is very much in the spirit of existing natural language generation (NLG) systems using overgeneration to handle multiple realisation possibilities for the same content, but this time we are using it for evaluating different *content*.

In this architecture, the inference system proposes sequences of logical formulae that follow from the ontology axioms (NLDI) and sends them to the realisation component. The realisation component selects the "best" realisation (given constraints on, for instance, sentence complexity) and this is evaluated to give a feedback score used to guide the generation of further alternatives via a best-first search mechanism.



Example: "What is an Electrode?"

Using a fuel cell ontology with 133 axioms, a relevance filter reduces the set of axioms to 31, including:

(A1) *Electrode* $\subseteq$ *Actuality*

(A2) *Electrode* $\subseteq \exists$ *contains. Catalyst*

(A3) *Electrode* $\subseteq$ ($\exists$ *contains. Support* $\cap \leq 1$ *contains.* $\top$)

(A4) *Domain(contains,*

FuelCell $\cup$ *MEA* $\cup$ *Electrode* $\cup$ *Catalyst)*

A successful sequence of search states is:

Electrode $\subseteq$ *Actuality*

(choose axiom (A1))

Electrode $\subseteq$ *Actuality*

Electrode $\subseteq$ *=1 contains. Catalyst*

(add axiom A2 with completed cardinality information)

Electrode $\subseteq$ *Actuality*

Electrode $\subseteq$ *=1 contains. (Catalyst* $\cap$ *Support)*

(aggregate with axiom (A3))

Electrode $\subseteq$ *Actuality*

Electrode $\subseteq$ *=1 contains. (Catalyst* $\cap$ *Support)*

Domain (contains,

FuelCell $\cup$ *MEA* $\cup$ *Electrode* $\cup$ *Catalyst)*

(add axiom (A4)) The final text is:

An Electrode is a kind of Actuality. An Electrode contains exactly one thing, which must be a Catalyst and a Support. Only something which is a fuelCell, a MEA, an Electrode or a Catalyst contains something.

NLG and Ontologies

Natural language generation has previously been connected to ontologies in two main ways. .irstly, following on from the ideas of "upper modelling" (Bateman 1990), ontologies have been used to help an NLG system organise its own knowledge and to aid portability of NLG systems. Secondly, NLG systems have taken domain knowledge expressed using an ontology, or even the ontology axioms themselves, as their inputs. Presenting ontologies in natural language, as motivated above, involves the second kind of connection between NLG and ontologies. Within this, unlike some previous work (Bontcheva and Wilks 2004) it involves presenting the ontology axioms rather than information about individuals which happens to be expressed using the terms of an ontology.

The use of natural language generation to express the definition of concepts in an ontology was pioneered in the GALEN-IN-USE project, which produced descriptions of surgical procedures in several European languages (Wagner et al 1999).

Bateman, J., "Upper modelling: organising knowledge for natural language processing", *Procs of the 5th International Workshop on NLG*, 1990.

Bontcheva, K. and Wilks, +., "Automatic report generation from ontologies: the MIAKT approach", *Procs of NLDB'04*, 2004.

Wagner, J., Rogers, J., Baud, R. and Scherrer, J-R., "Natural Language Generation of Surgical Procedures", *Medical Informatics* 53, 1999.