Motivation for "Ontology" in Parallel-Text Information Extraction

Mary McGee Wood and Shenghui Wang
Department of Computer Science, University of Manchester, UK
Email: {mary, wangs}@cs.man.ac.uk

Problems …

Legacy data in the mature descriptive sciences is often distributed across multiple text descriptions. For instance, in botany, a plant could be described in several text-based Floras.

- How to access these information resources in an easier way?
- How to correlate information from different sources?
- Do different sources agree with each other?
- Which is the best way to represent the information which is very complicated?

Solutions…

Ontology-based Information Extraction and Integration from parallel texts

- Extracting information from different texts separately referring to ontology
- Collecting results from different sources based on ontology
- Integrating information into an ontological knowledge base.

Motivation for Ontology

"Template" in traditional Information Extraction (IE) technology, developed within the MUC programme, is adequate when information needed is simple, flat and minimally hierarchical. However, extending IE to the complex, real-world domain of botany requires a major step up in the quantity and complexity of information to be extracted and represented. When our project [1,2] set out to acquire, in flexible electronic format, some of the wealth of legacy data locked in botanical descriptive texts (Floras), we rapidly reached the limits of the MUC template format for knowledge representation, finding it cumbersome and insufficiently expressive for our domain. We have therefore integrated the established GATE language processing system [3] with a range of tools for building and using ontologies, including Protege [4] for ontology development, and a Sesame [5] knowledge repository to store ontological data.

Ontology, compared to a template, has two advantages:

- Ontology represents better the scale and complexity of information in this domain
- Ontology offers a principled structure within which to merge the information extracted from parallel texts

Ontology-based Information Extraction

The system starts with a base upper ontology (developed and debugged in Protege, as Figure 1 shows), containing more than 70 object classes related to the botanical domain and around 20 properties that can hold between them. The output of the IE system, as Table 1 shows, is a copy of the base ontology populated with object instances, i.e. the heads and the features found by the textual analysis, and with property instances, i.e. relations between the object instances found in the text. The heads and the features identified by an IE sub-system, ontological gazetteer lookup which is linked with the base ontology, and instances generated for populating the ontology. The ontological gazetteer used in our system lists more than 2000 entries in 43 separate lists (each list being linked to a class in the base ontology).

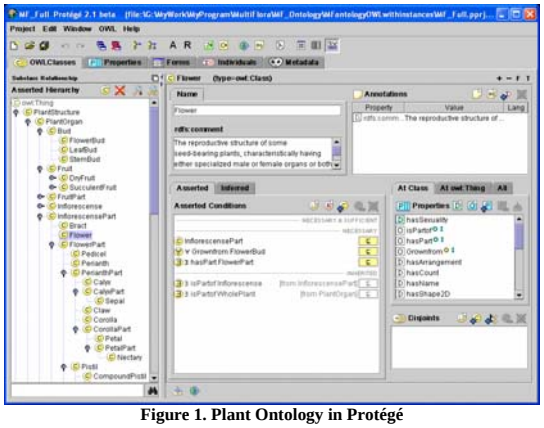

Figure 1. Plant Ontology in Protégé

Source Text:
Perennial herb with overwintering lf-rosettes from the short oblique to erect premorse stock up to 5 cm, rarely longer and more rhizome-like; roots white, rather fleshy, little branched.

System Output:

Head Class	Head	Property	Feature Class	Feature
Plant	herb	hasLifeform	Lifeform	Perennial
Leaf	Lf-rosettes	hasLifeform	Lifeform	overwintering
Stem	stock	hasRelativeProperty	RelativeProperty	short
Stem	stock	hasOrientation	Orientation	oblique to erect
Stem	stock	hasLength	Length	up to 5 cm
Stem	stock	hasRelativeProperty	RelativeProperty	more rhizome-like
Root	roots	hasColor	Color	white
Root	roots	hasShape3D	Shape3D	rather fleshy
Root	roots	hasShape3D	Shape3D	Little branched

Table 1. Output of IE referring to ontology

Ontology-based Information Integration

Integrating information extracted from such parallel texts compensates both for gaps in the texts themselves and for failures in processing. The results presented below demonstrate the value of this approach, and the role of ontology in implementing it.

Our IE system populates a single empty copy of the base ontology with information from all the contributing text sources. This gives us results like Figure 2:

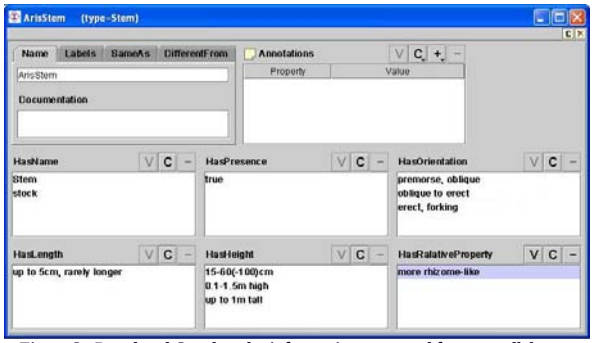

Figure 2. Populated Ontology by information extracted from parallel texts

Using a single resource for both processing of text and representation of results gives us a number of benefits, including consistency across parts of the system, and the modular localization of domain-specific knowledge. Although building such knowledge models is never trivial, having a clear and discrete place for them in one's system architecture is obviously an advantage.

Conclusions and prospects

We have argued that for many tasks (including, perhaps surprisingly, Information Retrieval in Biomedicine), simple template formats for information are adequate; and therefore preferable to the power and cost of an ontology. Where the scale and complexity of information to be handled is greater (as in our botanical domain), or the task requires more reasoning (as in our parallel-text processing), then the greater expressive power of an ontology is justified and valuable.

In the context of any discussion of emergent standards, we would plead for a standard, rigorous, meaningful definition of the term ``ontology", and for care, consistency, and respect in using either the term, or the thing itself.

Reference:

[1] Wood, M.M., S.J. Lydon, V. Tablan, D. Maynard, H. Cunningham. *Using parallel texts to improve recall in IE*. Recent Advances in Natural Language processing: Selected Paper from RANLP 2003, John Benjamins, Amsterdam/ Philadelphia (in press).
[2] Wood, M.M., S.J. Lydon, V. Tablan, D. Maynard, H. Cunningham. forthcoming. *Populating a Database from Parallel Texts using Ontology-based Information Extraction.* Proceedings OF NLDB '04.
[3] Cunningham, H., D. Maynard, K. Bontcheva, & V. Tablan. 2002. *GATE: A framework and graphical development environment for robust NLP tools and applications*. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia.
[4] J. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubézy, H. Eriksson, N. F. Noy, S. W. Tu *The Evolution of Protégé: An Environment for Knowledge-Based Systems Development.* 2002.
[5] J. Broekstra, A. Kampman, and F. Hanmelen.*Sesame: A Generic Architecture for Storing and Querying RDF*, International Semantic Web Conference 2002, Sardinia, Italy

Address for correspondence: Department of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL