# The PIA Project: Learning to Semantically Annotate Texts from an Ontology and XML-Instance Data

Nigel Collier
National Institute of
Informatics (NII)
National Center of Sciences,
2-1-2 Hitotsubashi
Chiyoda-ku, Tokyo 101-8430,
Japan
collier@nii.ac.jp

Koichi Takeuchi
National Institute of
Informatics (NII)
National Center of Sciences,
2-1-2 Hitotsubashi
Chiyoda-ku, Tokyo 101-8430,
Japan
koichi@nii.ac.jp

Keita Tsuji
National Institute of
Informatics (NII)
National Center of Sciences,
2-1-2 Hitotsubashi
Chiyoda-ku, Tokyo 101-8430,
Japan
keita@nii.ac.jp

## ABSTRACT

The development of the XML and RDF(S) standards offer a positive environment for machine learning to enable the automatic XML-annotation of texts that can encourage the extension of Semantic Web applications. After reviewing the current limitations of information extraction technology, specifically its lack of portability to new domains, we introduce the PIA project for automatically XML-annotating domain-based texts using example XML texts and an ontology for supervised training.

## 1. INTRODUCTION

PIA aims to develop a domain and language portable information extraction (IE) system. In contrast to other Web-based technologies such as information retrieval (IR) which are characterized by strong portability, no such system as yet exists for IE. We consider that the main factors which have prevented this are: (1) A focus within the IE community on general news-based IE, exemplified by systems that resulted from the message understanding conferences (MUCs) [6], and, (2) Despite recent moves towards machine learning for low level IE tasks such as named entity recognition there is still a strong reliance on large lexical resources such as term lists, and an emphasis on hand-built rules and patterns. The problem we see with this direction is that it promotes the development of rather inflexible IE systems that cannot easily be ported to new domains without substantial efforts to customize the system with domain-specific knowledge resources, e.g. the collection of domain dictionaries, writing domain-specific rules etc. Perhaps the greatest problem is that since there is no prior understanding between the IE system developer and the domain knowledge provider about the encoding of the knowledge that will be used to train the IE system, there is no guarantee that the type of knowledge that the system needs will be available in the new domain.

While the MUCs have made great advances in promoting the formalisation of IE tasks and evaluation, the MUC-style of IE technology provides a relatively sterile semantic environment. Semantics is limited to the border between syntax and semantics that occurs at the word and term level but largely ignores higher-level relations between the classes themselves and class-relations are ‘forced’ into disjoint relations as far as possible. The markup of text, while conforming to SGML, makes no use of an explicit ontology and relatively little use of meta-data.

Recent IE projects have looked beyond news to the molecular biology domain, e.g. [4] [5]. Some projects have implicitly incorporated simple taxonomies (is-a hierarchy) into the annotation guidelines for domain experts. To the best of our knowledge these projects still largely ignore explicit properties of classes and class relations that could be contained in the ontology and their potential contribution to automatic annotation. There seems to be great potential in this technical domain for incorporating ontologies into the learning model since a large amount of research has taken place on their development for gene-product databases such as SwissProt [1].

We believe then that with the advent of standards for the annotation of semantic content such as XML [3] for document structure, RDF [7] for defining objects and their relations, and RDFS [8] for defining the object model for describing RDF, that sources of domain knowledge will become widely available in electronic form and that these resources can and should be used for supervised training of a portable IE system which we call PIA-Core. Crucially these sources of knowledge will be available in a predictable format allowing PIA-Core to be rapidly deployed in a new domain.

## 2. PIA-CORE

We consider the W3C standardization process of XML and RDF(S) to offer a positive environment for machine learning of expert knowledge. Although ontologies in RDF are likely to emerge primarily as a result of (human) expert introspection we cannot expect that XML-instances of the defined concepts, such as technical terms, proper nouns, quantity and time expressions and their relations, will be annotated by experts for every document due to the high

*Semantic Web Workshop (SWWS)* 2001 Stanford University, USA

cost. This is one of the bottlenecks in the extension of Semantic Web [2] applications to the majority of documents that can be viewed on the Web and Intranets today. What is missing in the current focus on formalisation is a consideration about how the actual instantiation of the concepts defined in the ontologies will take place.

Actual semantic annotation of terminology and relations involves considerable time by domain experts and for this reason we believe it is worth investing in machine learning as a way to reliably replicate the capabilities of experts. This is the goal of PIA-Core. The scenario is that experts will develop a domain model in RDFS and a relatively small set of example annotated texts using an integrated XML and ontology editor. From this knowledge PIA will learn how to automatically XML-annotate unseen texts in the same domain. By focussing on domain-based learning we hope to make use of the ontology as a valuable knowledge resource and also to reduce the problems of ambiguity that developers of general IE systems must face.

In combination with robust domain-independent natural language processing (NLP) modules such as part of speech taggers, chunkers and shallow parsers, as well as general linguistic resources such as thesauri, PIA-Core will be used to XML-annotate texts that are consistent with those in the training set. We hope that PIA-Core can provide rapid acquisition of domain-knowledge and provide functionality that can be used at the heart of an IE system or within XML-tagging tools for computer-aided annotation. This can then serve as the basis for the deployment of 'smart' applications providing intelligent services that we hope to see emerge on the 2nd generation World Wide Web (Web), i.e. the 'Semantic Web'. The application we want to apply PIA-Core to is domain-based question-answering (e.g. [9]) in English and Japanese.

# 3. DISCUSSION

Some of the key questions that we need to consider are:

- How should we integrate the ontology into a statistically-based machine learning (ML) model? For example, how should we make use of concepts that appear in the ontology but don't appear in the training set? How can statistical evidence from sub-classes be shared through ontological relations to help overcome data-sparseness problems?

- XML is in some respects quite limited in its ability to represent complex object structure and relations as it is designed to encode serialization. We need to explore the limits of this representation for practical annotation of terminology and relations.

- Ontologies change over time - they may be revised, expanded or incorporated into other (shared) ontologies. How do we update the knowledge base that was extracted from the training set based on the original ontology? Should the model be retrained every time a change is made to the ontology? How can changes in the ontology be reflected automatically at the text markup level? E.g. the introduction of a new subclass.

- How will the issue of multi-linguality affect the design of ontologies in RDF for PIA?

# 4. CONCLUSION

We briefly presented a critical analysis of the current status of IE research and proposed a new project called PIA based on domain-based learning through XML-annotated texts and domain models described in RDF. We also considered some of the key research issues. From now we intend to implement PIA-Core and apply it to the task of question answering in technical domains such as molecular-biology.

# 5. ADDITIONAL AUTHORS

Gareth Jones (Department of Computer Science, University of Exeter, UK, email: gareth@dcs.exeter.ac.uk), Jun'ichi Fukumoto (Department of Computer Science, Ritsumeikan University, Japan, email: fukumoto@cs.ritsumei.ac.jp), Norihiro Ogata (Faculty of Language and Culture, Osaka University, Japan, email: ogata@lang.osaka-u.ac.jp), Chikashi Nobata (CRL, Japan, email: nova@crl.go.jp), Kinji Ono (National Institute of Informatics, Japan, email: ono@nii.ac.jp).

# 6. REFERENCES

[1] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Research*, 25:31–36, 1997.

[2] T. Berners-Lee. *Weaving the Web*. Harper, San Francisco, 1999.

[3] Namespaces in xml, world wide web consortium recommendation. http://www.w3.org/xml/TR/REC-XML, 14th January 1998.

[4] N. Collier, H. Park, N. Ogata, Y. Tateishi, C. Nobata, T. Ohta, T. Sekimizu, H. Imai, and J. Tsujii. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. In *Proceedings of the Annual Meeting of the European chapter of the Association for Computational Linguistics (EACL'99)*, June 1999.

[5] M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the 7th International Conference on Intelligent Systemps for Molecular Biology (ISMB-99)*, Heidelburg, Germany, August 6–10 1999.

[6] DARPA. *Proceedings of the Sixth Message Understanding Conference(MUC-6)*, Columbia, MD, USA, November 1995. Morgan Kaufmann.

[7] Resource description framework (rdf) model and syntax specification, world wide web consortium recommendation. http://www.w3.org/xml/TR/REC-rdf-syntax, 22nd February 1999.

[8] Resource description framework (rdf) schema specification 1.0, w3c candidate recommendation. http://www.w3.org/xml/TR/2000/CR-rdf-schema-20000327, 27th March 2000.

[9] R. Srihari and W. Li. Information extraction supported question answering. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8), Gaithersburg, Maryland, USA*. National Institute of Standards and Technology (NIST), November 17–19 1999.