PIA-Core: Semantic Annotation through Example-based Learning

Nigel Collier* **and Koichi Takeuchi***

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

Abstract

This paper summarizes the aims and scope of the PIA (Portable Information Access) project's PIA-Core system for automatic annotation of documents on the Semantic Web, i.e. the next generation World Wide Web. The focus of the project is to develop a portable information extraction system that can be easily adapted to new domains. PIA has its foundations on three resources: the PIA-Core information extraction module, application modules and PIA guidelines for ensuring consistent annotation. We are currently developing PIA-Core based on advanced machines learning methods to automatically annotate documents with terminology, names, temporal and quantity expressions etc. using examples of annotated documents.

1. Introduction

PIA aims to develop a domain and language portable information extraction (IE) system. Although advanced IE systems do exist, 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. Perhaps 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) (MUC, 1995), and, (2) Despite progress 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 *a priori* 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. We believe that the Semantic Web offers an opportunity to solve some of these problems.

2. Machine Learning on the Semantic Web

The Semantic Web model (Berners-Lee et al., 1999), now being proposed by the W3C (World Wide Web Consortium) as the next generation Web raises many exciting possibilities. For example, that we can annotate instances of classes and relations according to an ontology written in in either RDF Schema (Brickley and Guha, 2000) or DAML+OIL (Hendler and McGuiness, 2000) and then build software applications to gather and reason with this knowledge using inference engines built on logic. Ontologies, which may be considered to be "a specification of a conceptualization" (Gruber, 1993) are used for knowledge sharing and re-use and are the basis of this model. Intelligent applications such as being able to find the answer to a question in a document collection, electronic shopping, making appointments using agents (*information brokering*), as well as 'smart' browsing of documents can then become a reality. The Semantic Web should also contribute to enabling language transparency of documents.

The majority of information on the Web, estimated at about 70%, is in the form of free-texts. However, due to the very high cost and time required we cannot expect that instances of the concepts defined in the ontologies will be marked up by experts in every text. It is also difficult to ensure quality of annotation: both in terms of consistency and coverage. This is one of the bottlenecks in the extension of Semantic Web applications to the majority of documents that can be viewed on the Web today. What is missing in the current focus on formalization is a consideration about how the actual instantiation of the concepts defined in the ontologies will take place. We believe therefore that it is worth exploring machine learning as a way to reliably replicate the capabilities of experts. This is the goal of PIA-Core.

Our expectation is that with the advent of standards for the annotation of semantic content on the Web such as XML (Bray et al., 2000) for document structure, RDF for defining objects and their relations, and RDF(S) (RDF Schema) for defining basic ontological modelling primitives on top of RDF, that sources of domain knowledge will become widely available in electronic form and that these resources 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 new domains. In this respect the requirement of IE for structured knowledge and of the Semantic Web for instantiation can be viewed as complementary.

3. Annotation of NE+ Expressions

PIA-Core's basic motivation is similar to that of previous IE systems, i.e. the extraction of prototypical facts rather than full understanding of a text. Examples could include take-overs or mergers between companies, signal transductions between genetic products, reporting of macro-economic data or sports results. PIA-Core extends

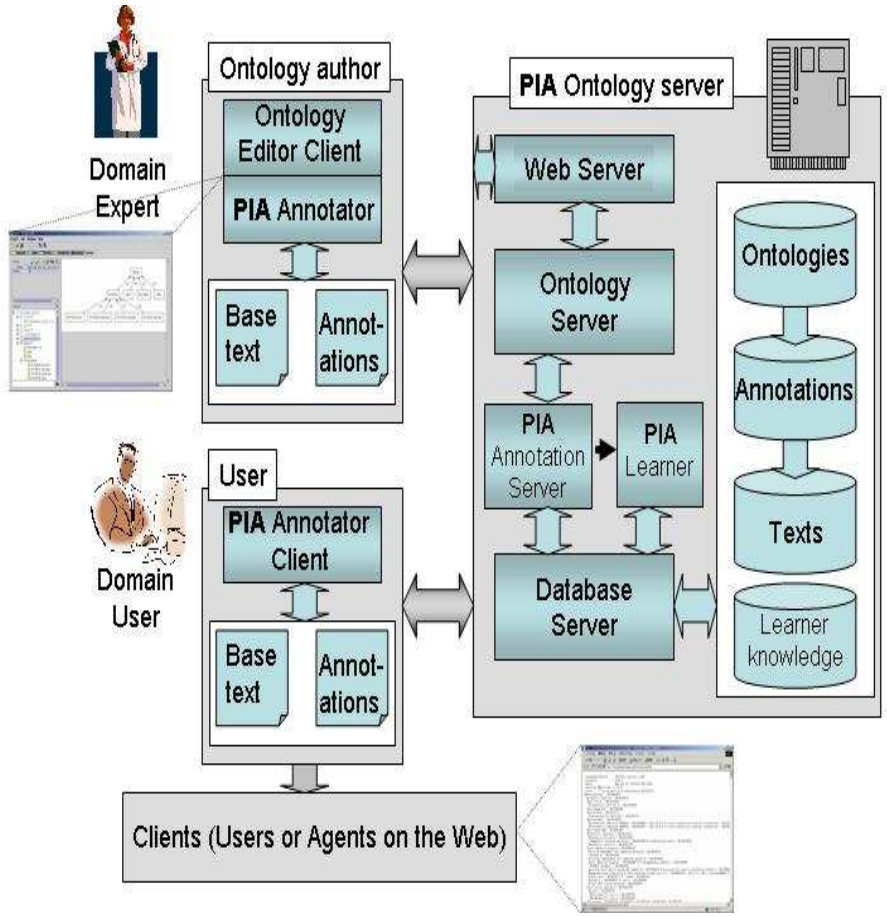

Figure 1: Logical overview of the PIA system. PIA-Core is shown here as the annotation server and learner.

previous IE systems though by capturing what we call named entity plus (NE+) expressions, i.e. types as well as individuals, where the types are members of an explicitly defi ned ontology.

A concise summary of the guidelines for NE+ annotation is provided in (Collier et al., 2002) and the full guidelines for NE+ annotation will be published as an NII technical report. Briefly, candidates for inclusion in NE+ are:

- proper nouns, e.g. names of people, places
- temporal expressions, e.g. days of the week, dates
- quantity defi nitions, e.g. names of monetary values, names of stock market indices
- terminological expressions
- certain expressions that share identity with the above.

We briefly comment here on the aspects of NE+ which will require extensions to the traditional NE models used by machine learning-based IE systems.

The NE+ guidelines extend markup conventions from so-called traditional named entity expressions, i.e. the annotation of individuals such as people, places or organizations to annotation of types in the form of technical terms. NE+ expressions have a much wider range of surface variant forms compared to NE expressions and so the guidelines provide rules for converting surface forms to what we call 'conventional forms'. Rules cover various forms of transformation resulting from fi ve main areas: graphical variations, inflectional variations, shallow syntactic variations such as conjunction, semantic variations including issues of granularity according to the ontology, and discourse variations such as the use of abbreviations, aliases, pronouns and defi nite descriptions. In addition to conventional forms the guidelines make the explicit connection between an NE+ as an instance of a class in an ontology as well as the linkage from the NE+ to its position in the base document. The guidelines provide an RDF Schema that defi nes the name space for annotations.

In PIA-Core we are considering a number of issues for machine learning of NE+ expressions which extend previous NE technology. These include:

- The need for rich feature sets to support annotation of technical terms - and appropriate models to support this;
- The need to consider how to model nested semantic structures which are supported in the NE+ guidelines;
- The need to consider taxonomic relations between classes in the ontology and how this can benefi t the model;

4. Current Status

The work on PIA-Core is still 'work in progress'. We report here briefly on a number of component technologies.

As shown in Figure (1), the scenario is that experts will develop a domain ontology and a relatively small set of example annotated texts. To aid in annotation we will soon release annotation guidelines (English and Japanese) for the NE+ task. To support development of annotation data according to the guidelines we have released the first version of an annotation client called PAT which supports the NE+ guidelines. PAT v1.0 is a Java plug-in for the Protégé-2000 ontology editor (Noy et al., 2001). A screen shot of the tool is shown in Figure 2. The first version supports linkage between annotations and the base text using a simple byte-start and byte-end pointer. In the second version which we are now making this will be upgraded to an XPointer (De Rose et al., 2000) notation, and the tool itself will be optimized for speed and memory management in C++.

From this knowledge PIA-Core will learn how to automatically annotate new texts in the same domain. Domain users can then choose an ontology from the ontology server and have their documents annotated by the annotation server so that they are consistent with the example annotations. Our current application domains are molecular biology and news; we aim to develop the system initially for both English and Japanese with additional guideline support for Thai and Arabic.

We are now in the process of developing a new test collection in the domain of molecular biology based on the annotation of 9 top-level classes of genes and gene products for about 100 EMBO Journal articles. Each article is annotated by an expert qualified at Ph.D. level according to an early draft of the NE+ guidelines and using the PAT annotation tool. On average each text provides about 1000 annotations of NEs. The resulting corpus will provide a rich test environment for our work in domain-based NE annotation.

We are now investigating SVMs (Support Vector Machine) (Cristianini and Shawe-Taylor, 2000) and HMMs (Hidden Markov Model) (Rabiner and Juang, 1986) for the annotation task that combine the knowledge available in the ontology with linguistically motivated features available from robust natural language processing tools such as a shallow parser. Currently we have implemented NE models for learning 'flat' rather than nested semantic structures, and we are exploring how to incorporate very rich features sets into the model.

5. Conclusion

In PIA-Core we are developing a set of tools for deriving semantic content from Web-based documents according to example-based learning. The types of semantic content we are focussing on includes terminology, names, temporal and value expressions as well as coreference and relations.

If we can achieve our goal then we hope that PIA-Core can provide a domain portable information extraction system that contributes to the increase of knowledge available to intelligent computer applications and users on the Semantic Web.

6. References

T. Berners-Lee, M. Fischetti, and M. Dertouzos. 1999. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*. Harper, San Francisco, September. ISBN: 0062515861.

T. Bray, J. Paoli, C. Sperberg-McQueen, and Maler, E. 2000. Extensible markup language (xml) 1.0 (second edition). http://www.w3.org/TR/2000/REC-xml-20001006.

D. Brickley and R. V. Guha. 2000. Resource Description Framework (RDF) schema specification 1.0, W3C candidate recommendation. http://www.w3.org/TR/2000/CR-rdf-schema-20000327, 27th March.

N. Collier, J. Fukumoto, K. Takeuchi, N. Ogata, C. Nobata, and K. Tsuji. 2002. Progress on multi-lingual named entity annotation guidelines using DAML+OIL. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'2002), Las Palmas, Spain*, May 7–11.

N. Cristianini and J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, England; ISBN 0521780195.

S. De Rose, E. Maler, and Daniel, R. (eds). 2000. Xml pointer language (xpointer) version 1.0, w3c candidate recommendation, 11th september 2001. http://www.w3.org/TR/xptr.

T. R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 6(2):199–221.

J. Hendler and D. L. McGuiness. 2000. The DARPA Agent Markup Language. *IEEE Intelligent Systems Journal*, 16(6):63–73, Jan./Feb.

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

N. F. Noy, M. Sintek, S. Decker, M. Crubezy, R. W. Fergerson, and M. A. Musen. 2001. Creating semantic web contents with Protégé-2000. *IEEE Intelligent Systems*, 16(2):60–71.

L. Rabiner and B. Juang. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine*, pages 4–16, January.

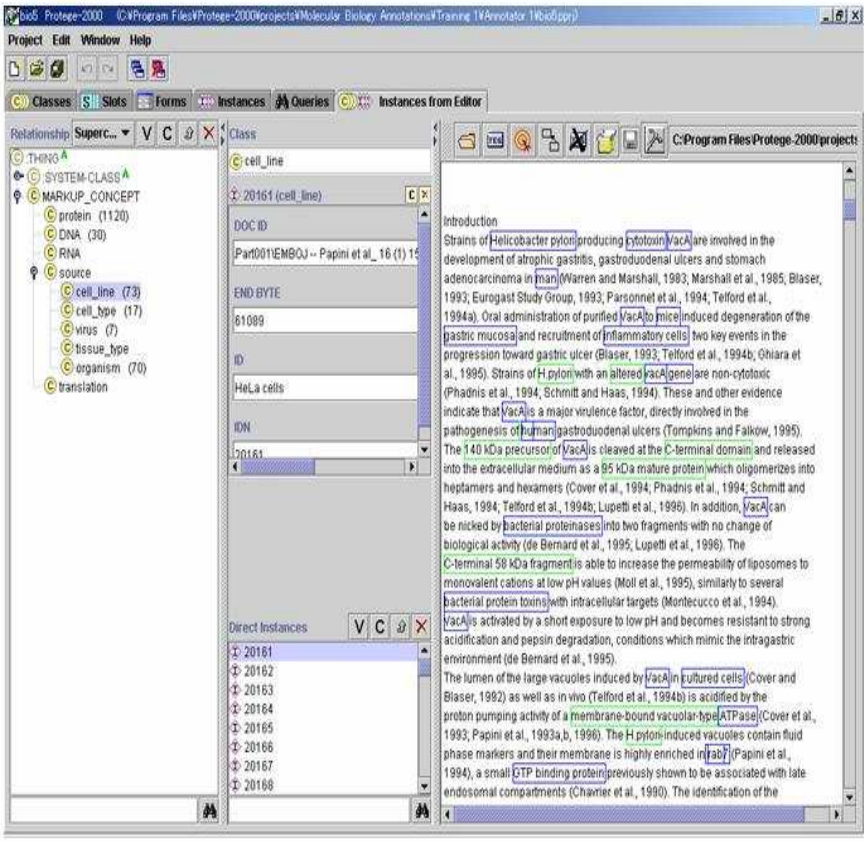

Figure 2: A screen shot of the PIA Annotation Tool (PAT) version 1.0 Java plug-in for Protégé-2000. NE+ expressions are shown in boxes in the text pane on the right, the ontology is shown in the left pane and instance information is shown in the middle.