Managing the semantics of coreference relations with Open Ontology Forge

Ai Kawazoe   Asanobu Kitamoto   Nigel Collier

National Institute of Informatics
2-1-2 Hitotsubashi Chiyoda-ku
Tokyo 101-8430 JAPAN
{zoeai, kitamoto, collier}@nii.ac.jp

Abstract. In this paper, we will discuss managing the semantics of "coreference relations" in the framework of the Semantic Web. In the Semantic Web context coreference is important in integrating many kinds of information sources (of various linguistic forms, images, etc.) and helping users to share such information. In this paper we propose a knowledge model for describing the semantics of co-referential identity relations on *annotations*, and introduce the Open Ontology Forge (OOF) software to support users to manually annotate coreference in texts and between images and texts.

1 Introduction

In this paper, we will discuss how to manage the semantics of coreference by human experts. In the domain of natural language processing, coreference has been a key problem for computers to understand the meaning of natural language texts through anaphoric expressions that require disambiguation, and accurate identification of coreference makes it possible for computers to maximize the amount of useful information. In the Semantic Web context, "coreference" play a role not only in augmenting information, but also in integrating information sources which refer to the same class instance and helping users to share the information. So far, the Semantic Web initiative [1] has enabled RDF [5] and OWL [7] to become a common meta-data standard for sharing knowledge on the World Wide Web, and this allows for the explicit description of concepts, properties and relations in an *ontology*. Instances of concepts in an ontology appear in documents in many different forms (ex. various linguistic forms, images, etc.) and in order to integrate information it is necessary to manage the coreference relations between such surface forms.

In this paper we outline a knowledge model for describing the semantics of coreference on *annotations* implemented within Open Ontology Forge (OOF), software to support users for annotating coreference relations by hand. The coreference cases which we aim to deal with in this work are taken from the domain of molecular biology and include 1) coreference among linguistic expressions including anaphoric relations (ex. Interleukin-2…it) and term variations (ex. Interleukin-2 vs. IL-2), 2) multimodal coreference among texts and images (ex. a biomedical image of a cell and description of the cell in a figure legend), and 3) cross-document coreference.

2 Representing Coreference

In the OOF knowledge model, we represent coreference relations as illustrated in Fig.1. In this model, co-referring expressions are related to what we call a "coreference pool" which is linked to an ontological class.

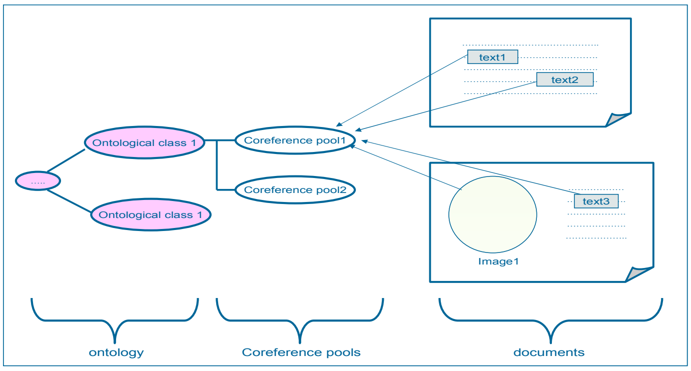
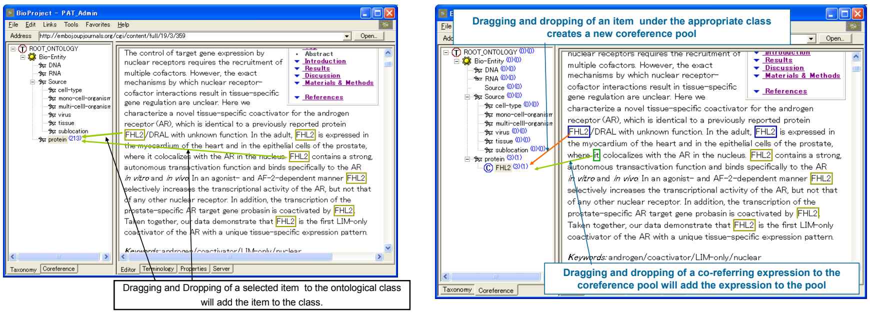


Fig. 1. Overview of the coreference annotation model.

The important feature of the model here is that the co-referring expressions in a coreference pool have the same status (unlike "coreference sets" in Conceptual Graph) and there is no hierarchical relation among them. Each of the expressions, regardless of their type, are independently related to a coreference pool. From a practical point of view, we can say that this style of annotation is one of the easiest ways to represent cross-modal and cross-document coreference. Also in intra-text coreference annotation, it will reduce a lot of user efforts and make no assumption of linguistic training, especially in that it does not require specifying antecedents for pronouns or canonical forms for names, unlike other schemes such as MUC [4].

3 OOF annotation support tool

We have developed the OOF software to support users in annotating coreference relations by hand. The main features of the software are 1) an integrated function for ontology creation and text/image annotation, and 2) a user interface which realizes an easy way of annotation for cross-modal items. OOF has a full Web-browser view of a html document, along with a window showing the ontology and coreference pools. Users can create the class hierarchy by expanding the root class and defining new class names. The software allows for two modes of instance capture: the named entity annotation mode and the coreference annotation mode. In both, users can make text annotation by dragging and dropping the selected part of text (Fig.2) to the class in the taxonomy. For image annotation, OOF has an SVG editor window for selecting a part of image and editing properties as in Fig. 3. The selected part of the image can be linked to an ontological class or a coreference pool in the same manner as texts: a simple drag-and-drop fashion.



The control of target gene expression by nuclear receptors requires the recruitment of multiple cofactors. However, the exact mechanisms by which nuclear receptor–cofactor interactions result in tissue-specific gene regulation are unclear. Here we characterize a novel tissue-specific coactivator for the androgen receptor (AR), which is identical to a previously reported protein FHL2/DRAL with unknown function. In the adult, FHL2 is expressed in the myocardium of the heart and in the epithelial cells of the prostate, where it colocalizes with the AR in the nucleus. FHL2 contains a strong, autonomous transactivation function and binds specifically to the AR *in vitro* and *in vivo*. In an agonist- and AF-2-dependent manner FHL2 selectively increases the transcriptional activity of the AR, but not that of any other nuclear receptor. In addition, the transcription of the prostate-specific AR target gene probasin is coactivated by FHL2. Taken together, our data demonstrate that FHL2 is the first LIM-only coactivator of the AR with a unique tissue-specific expression pattern.

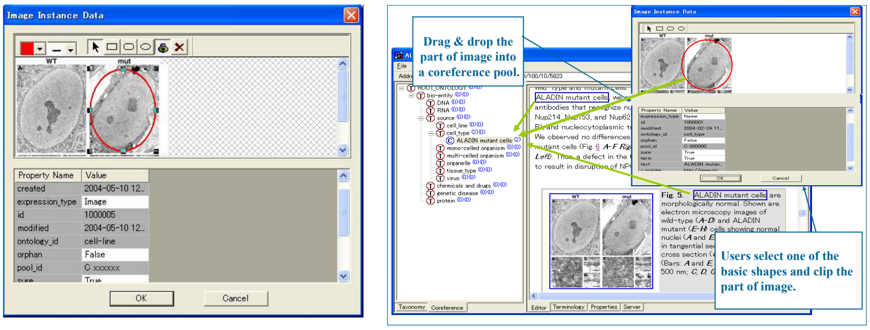
Keywords: androgen/coactivator/LIM-only/nuclear

Dragging and Dropping of a selected item to the ontological class will add the item to the class.

Dragging and dropping of an item under the appropriate class creates a new coreference pool

Dragging and dropping of a co-referring expression to the coreference pool will add the expression to the pool

Fig. 2. (left) Named entity annotation and (right) coreference annotation with OOF.



Property Name	Value
created	2004-05-10 12...
expression_type	Image
id	1000005
modified	2004-05-10 12...
ontology_id	cell-line
orphan	False
pool_id	C xxxxxx
sure	True

Property Name	Value
expression_type	Name
id	1000001
modified	2004-02-24 11...
ontology_id	cell_type
orphan	False
pool_id	C 000000
sure	True
term	True
text	ALADIN mutan...

Fig. 5. ALADIN mutant cells are morphologically normal. Shown are electron microscopy images of wild-type (*A–D*) and ALADIN mutant (*E–H*) cells showing normal nuclei (*A* and *E*

Drag & drop the part of image into a coreference pool.

Users select one of the basic shapes and clip the part of image.

Fig. 3. (left) A window for image editing. (right) Annotation of text-image coreference

The OOF knowledge model has several similarities to other ontology editors such as Protégé-2000 [6] and OntoEdit [8]. What distinguishes OOF is its focus on providing support for content annotation. An annotation in OOF is regarded as an *instance* with a linkage to the document and tracking information about the annotator, recorded by the pre-defined properties including (1-3) below, whose values are automatically assigned by the software. Annotations are grouped within 'coreference pools' via (4), and (5) and (6) to record characteristics of instances.

1. *XPointer* takes an XPointer value [2] and relates an annotation to the resource
2. *author* is a name of the author of the annotation
3. *ontology_id* relates an annotation to an ontological class
4. *pool_id* takes an ID of the *coreference pool* to which the item belong
5. *expression_type* specifies the subtype of the instance (*name, pronoun, image, etc.*)
6. *svg* records a description of the annotated part of image in SVG [3].

4 Discussion

We have conducted two kinds of annotation case studies using OOF: 1) annotation of biomedical articles (text only), and 2) annotation of documents which include images of Buddhist statues in Dunhuang. We had two biologists and some experts in the cultural heritage domain as annotators, who are not experts in NLP. In the former experiment, the annotators seemed to have a good understanding of the notion of coreference pool, and they made coreference annotations in the way we have intended. However, since the current version of OOF does not have semi-automatic annotation function for coreference, some coreference relations were left unannotated. The latter experiment have revealed some limitations of the current OOF knowledge model. For example, when annotating an image of a Buddhist statue, users sometimes want the same image to be both under a class of styles (ex. Ghandara_style) and under a class of motifs (ex. Bodhisattva), but OOF does not allow such annotation. Further, relations such as a part-of relation seem necessary, where we are describing the relationship between a Buddhist statue and its halo. We should reconsider how to manage these situations with OOF.

5 Concluding remarks

Open Ontology Forge was released in February 2004 and available freely for download from http://research.nii.ac.jp/~collier/resources/OOF/index.htm. We plan to release a new version of OOF in December 2004 with several functions including multi-document annotation.

References

1. Berners-Lee, T., Fischetti, M., and Dertouzos, M. 1999. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*. Harper, San Francisco, September.
2. DeRose, S., Maler, E., and Daniel, R. eds. 2001. XML Pointer Language (XPointer) Version 1.0. W3C can-didate recommendation, 11th September
3. Ferraiolo, J., Fujisawa, J., Jackson, D. 2003. Scalable Vector Graphics (SVG) 1.1 Specification, *W3C Recommendation* 14 January 2003. (http://www.w3.org/TR/SVG11/)
4. Hirschman, L., and Chinchor, N. 1997. MUC-7 Coreference Task Definition, Version 3.0. In *Proceedings of the Seventh Message Understanding Conference.*
5. Lassila, O., and Swick, R. eds. 1999. Resource Description Framework (RDF) Model and Syntax Specification. Recommendation, W3C, February.
6. Noy, N. F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R. W., and Musen, M. A. 2001. Creating semantic web contents with Prot´eg´e-2000. In *IEEE Intelligent Systems*, 16(2):60–71.
7. Smith, M. K., Chris Welty, C., McGuinness, D. L. (eds.) OWL Web Ontology Language Guide, W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-guide/
8. Sure, Y., M. Erdmann, J. Angele, S. Staab, S. Studer, and D. Wenke. 2002. OntoEdit: Collaborative ontology engineering for the semantic web. In I Horrocks and J. Hendler, editors, *The Semantic Web - ISWC 2002:First International Semantic Web Conference,Sardinia, Italy*, Lecture notes in Computer Science. Springer-Verlag,