ANNOTATION OF COREFERENCE RELATIONS AMONG LINGUISTIC EXPRESSIONS AND IMAGES IN BIOLOGICAL ARTICLES

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 propose an annotation scheme which can be used not only for annotating coreference relations between linguistic expressions, but also those among linguistic expressions and images, in scientific texts such as biomedical articles. Images in biomedical domain often contain important information for analyses and diagnoses, and we consider that linking images to textual descriptions of their semantic contents in terms of coreference relations is useful for multimodal access to the information. We present our annotation scheme and the concept of a "coreference pool," which plays a central role in the scheme. We also introduce a support tool for text annotation named Open Ontology Forge which we have already developed, and additional functions for the software to cover image annotations (ImageOF) which is now being developed.

1. Introduction

In this paper, we propose an annotation scheme which can be used not only for annotating coreference relations between linguistic expressions, but also those among linguistic expressions and images, in scientific texts such as biomedical articles.

"Coreference" is a relation among expressions which refer to the same thing, for example, "Gates" and "his" in the sentence "Gates grew up in Seattle with his two sisters." Coreference relations are important for tasks such as information extraction, since accurate identification of such relations in texts makes it possible for computers to maximize the amount of useful information they can understand. Annotation schemes for coreference and coreference resolution techniques based on the annotations are developed by many IE researchers. For example, the IE systems for the biomedical literature developed by (Castaño et al., 2002) and (Hahn et al., 2002) have their strength in the coreference resolution module.

Some previous studies have worked on coreference relations among objects of different media, including ones between texts and images. For example, (André and Rist, 1994) presents a model of referring by texts and images, and (van Deemter, 1998) proposes semantic representations for multimedia coreference taking "representationalist" approach. (Yamada and Nakagawa, 2002) proposes a method to identify coreference relation between names and human faces in photo journals.

Annotating such cross-media coreference relations in an appropriate way is highly desirable for multimodal access to the information in the domain of biomedicine, which is currently the main target of the IE technology. In this domain, where huge amounts of image data are produced every day as well as textual data, many efforts are made in tasks such as constructing databases of images accessible to researchers (ex. BioImage database project: Shotton, 2003), and retrieving image data automatically (Liu et al., 2004, among others.) Biomedical images constitute an essential component for analyses and diagnoses, and they often illustrates important facts and discoveries more effectively than textual descriptions. However, it is also true that they still needs to be linked to natural language descriptions of their semantic contents in some way, in order to be accessible by human researchers. We consider that the link between images and texts can be realized in terms of multimedia coreference.

In the rest of this paper we present our scheme to annotate coreference between images and texts in machine readable way. In Section 2, we will introduce a notion of a "coreference pool" which plays an important role in our annotation scheme for multimodal coreference. In Section 3, we will provide an overview of the metadata scheme, and in Section 4 we will introduce a support tool for text annotation named *Open Ontology Forge* which we have already developed, and additional functions for the software to cover image annotations named *ImageOF* which is now being developed.

2. Coreference pools

Although annotation schemes for coreference relations are proposed in many works including MUC-7 (Hirschman and Chinchor, 1997) among others, they are mainly for coreference relations among linguistic expressions, and not suitable for annotating coreference among multimodal objects. This is mainly because they describe coreference relations as a dependency relation between parts of texts and annotate them in a manner just specifying *antecedents* of coreference expressions, not referring to the identity of their referents. As (van Deemter, 1998) points out, it is not obvious that coreference relations among images and those between images and linguistic expressions can be properly captured as such a "dependency" relation involving the notion of antecedent, since distinction between anaphor and antecedent is not clear in multimedia coreference relations.

In our annotation scheme, "coreference pool" is a key notion, which makes multimodal coreference annotation easier and simpler. Coreference pools are sets of items (linguistic expressions/images/possibly other modes of expressions) which refer to the same object in the world. Since all of the items in the same pool have the same *symmetric* status without any dependencies to each other,

our coreference annotation scheme only require annotators to create a coreference pool and put every occurrence of co-referential items to the pool, without specifying which is the *antecedent* of which item as in many other existing schemes. Each coreference pool can contain both linguistic expressions (named entities and others) and images, and in this view it is perfectly natural to extend annotation to other modal (ex. sound). This way of "symmetric" annotation using coreference pools has some practical advantages, for example:

- It reduces a lot of efforts by annotators who do not have linguistic background, and avoids inconsistency in annotation
- Inter-text coreference can be easily captured, without specifying one "most conventional form" among co-referential expressions. This will avoid confusions and inconsistencies in choosing conventional forms among annotators.

An overview of our coreference annotation scheme is given in Figure 1. As shown in Figure 1, coreference pools are linked to ontology.

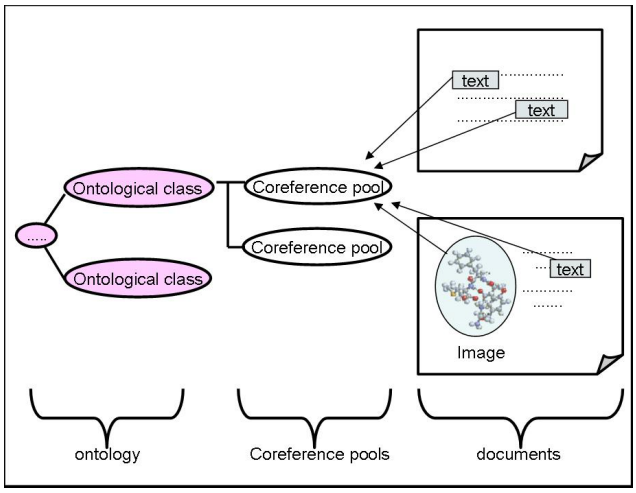

Figure 1: An overview of the annotation scheme

This way of coreference annotation also has a theoretical motivation, see (Kawazoe and Collier, 2003) for details.

3. Metadata schemes

Now we will introduce our metadata scheme. The central notion of our scheme is *ontology*, and *annotation* is its instance, which is defined in RDF Schema (Lassila et al., 1999) as a name space and held on the server at a URI. Ontologies are created by domain experts of biology, and basically they are free to create any classes that help define knowledge in their domain according to the limits of RDF Schema.

Each coreference pool is linked to an appropriate ontological class. Annotations of coreference expressions (including named entities) and relevant (parts of) images are saved in a file separated from the original text. Linkage between annotation and texts are automatically inserted as XPointer values (DeRose et al., 2001). We briefly now describe some of the metadata:

- *XPointer* relates an annotation to the resource to which the annotation applies and takes on an XPointer value
- *ontology_id* relates an annotation to the ontology and class to which the annotation applies
- *pool_id* is an ID of the *coreference pool* to which the item belong, and this is used to indicating coreference relations between annotations
- *term* takes an Boolean value and indicates if the item is a terminology or not
- *expression_type* specifies the subtype of the item (ex. *name, alias, pronoun, definite description, image,* etc.)
- *created* records the time when the annotation is created.
- *modified* records the latest time when the annotation is modified.
- *author* is a name of the author of the annotation
- *comment* is used when the annotator wants to leave some comments for others
- *sure* takes an Boolean value and indicates if the annotator is sure about his/her annotation or not. This information is for quality control of the annotation and can be used in post-annotation processing.
- *text* shows the annotated part of text literally. In the case of image annotation, this property takes a description of the annotated part of image in SVG.

In addition to these annotation classes, annotation of an image involves specification of relevant part of the image in terms of Scalable Vector Graphics: annotators can clip the parts of images in rectangles, circles, ellipses and other basic shapes available in SVG. This is then converted into an intermediate XML-like notation for storing in the properties of the captured instance.

An example of the annotation of a part of text and a part of image in XML is shown below. Notice that the annotated items have the same *pool_id* value, which shows that they are co-referential.

```
<annotation>
  <item>
    <properties>
      <author>zoeai</author>
      <created>2003-12-02 14:06:37</created>
      <expression_type>Name</expression_type>
      <id>1000000</id>
      <modified>2003-12-02 14:06:37</modified>
      <ontology_id>cell-type</ontology_id>
      <pool_id>C 000000</pool_id>
      <sure>True</sure>
      <term>True</term>
      <text>ALADIN mutant cells</text>
      <XPointer>http://www.pnas.org/cgi/content/
      full/100/10/5823#xpointer(string-
      range(//*,'ALADIN mutant
      cells')[2])</XPointer>
    </properties>
  </item>
  <item>
    <properties>
      <author>zoeai</author>
      <created>2003-12-02 14:06:37</created>
      <expression_type>Image</expression_type>
      <id>1000000</id>
      <modified>2003-12-02 14:06:37</modified>
      <ontology_id>cell-type</ontology_id>
      <pool_id>C 000000</pool_id>
      <sure>True</sure>
```

```
        <term>True</term>
        <text><svg
        xmlns="http://www.w3.org/2000/svg"
        version="1.1"><rect x="1" y="1"
        width="1198" height="398" /><image x="200"
        y="200" width="100px" height="100px"
        xlink:href="myimage.png"/></svg></text>
        <XPointer>http://www.pnas.org/cgi/content/
        full/100/10/5823(((/img[1])</XPointer>
      </properties>
    </item>
</annotation>
```

4. Open Ontology Forge and ImageOF

As a support tool for text annotation by human annotators, we have developed Open Ontology Forge (OOF), which provides a convenient environment for annotation of coreference relations. We are now developing a new annotation software based on OOF which has extended functions to annotate parts of images, named ImageOF. In this section we will briefly introduce some features of these tools.

OOF is a software to support ontology creation and text annotation combined to the ontology by human experts. OOF has many useful features that enhance functions found in previous ontology editors such as Protégé-2000 (Noy et. al, 2000). As shown in the screenshot of Figure 2, it has a full Web-browser view of the Web page including images, along with the window which shows created ontology and coreference pools under ontological classes.

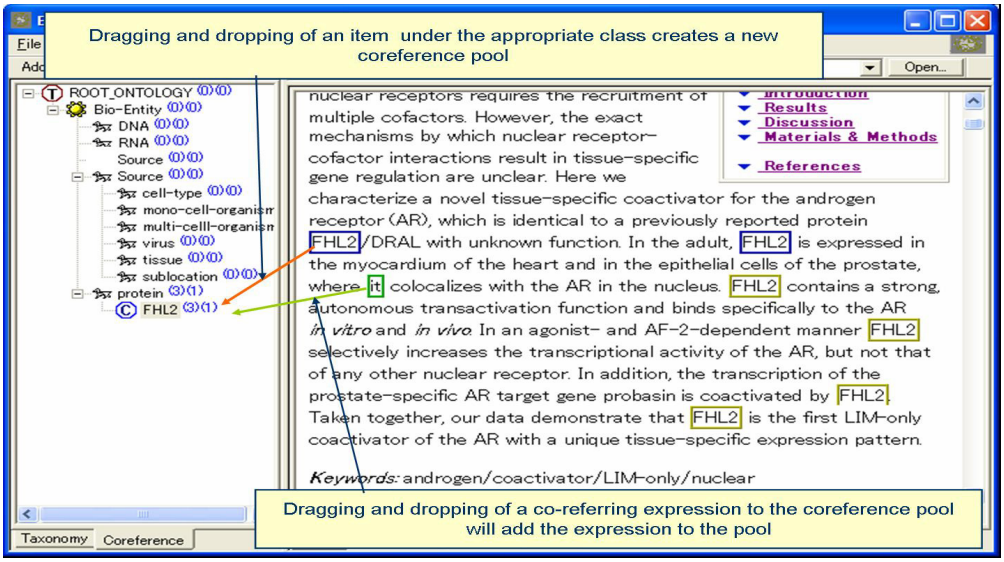

Figure 2: A screenshot of Open Ontology Forge which shows the coreference annotation procedure.

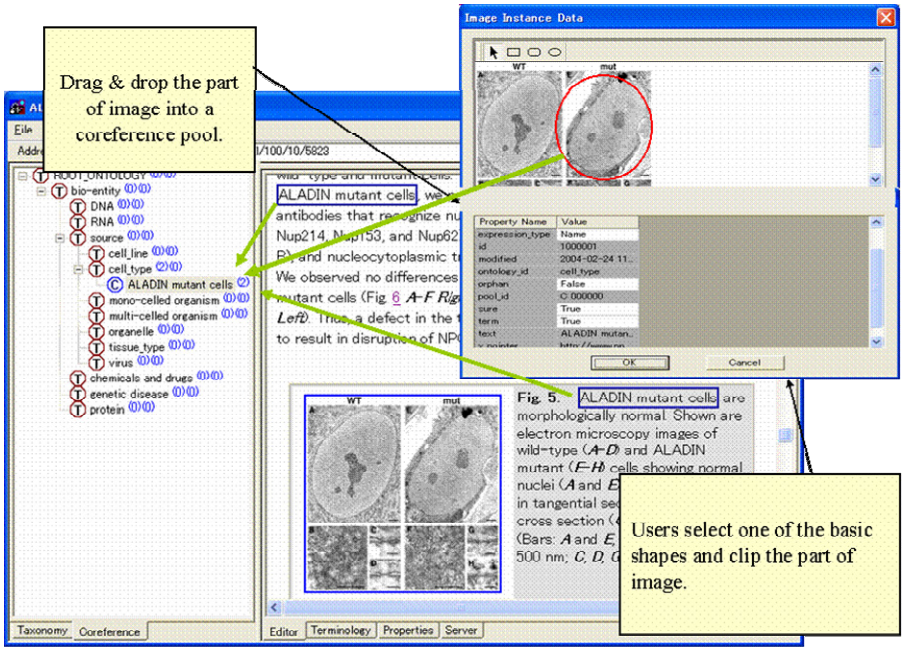

Figure 3: Design of the planned image annotation procedure with ImageOF

It has a two-modes of text annotations: named entity annotation and coreference annotation, and both of these annotations can be done by a semi-automatic, easy drag-and-drop manner. It also has automatic checking of consistency of properties among the members of coreference pool by unification. Users can export the result of the annotation to in-text XML format, and also can save ontology in RDF.

ImageOF is designed to realize the consistent procedure for linking (parts of) images with ontological classes and coreference pools. It will include a drawing window to clip a part of image in basic shapes available in SVG, an image visualization component for viewing and searching for annotated parts of images, and a drag-and-drop annotation procedure for images, as shown in Figure 3.

5. Concluding Remarks

Finally we briefly mention the PIA project (Collier et. al., 2003[1]) and its current status. This project aims 1) to provide an environment where domain experts can cooperatively create domain ontology on-line, and 2) to construct automatic text annotation system which makes annotations based on the domain ontology to unseen texts. We are now constructing annotated corpora for training from biological papers from publicly available source, using Open Ontology Forge.

Current version of OOF and user's manual are downloadable from:

http://research.nii.ac.jp/~collier/OOF/index.htm

We plan on releasing ImageOF in summer 2004 after user testing is completed.

References

André, E., and Rist, T. 1994. Referring to World Objects with Text and Pictures.In *Proceedings of COLING* 1994, Kyoto, Japan.

Castaño, R., Zhang, J., and Pustejovsky, J. 2002. Anaphora Resolution in Biomedical Literature. International Symposium on Reference Resolution for Natural Language Processing, Alicante, Spain.

Collier, N., Takeuchi, K., Kawazoe, A., Mullen, T., and Wattarujeekrit, T. 2003. A Framework for Integrating Deep and Shallow Semantic Structures in Text Mining. In *Proceedings of the Seventh International Conference on Knowledge-based Intelligent Information and Engineering Systems (KES'2003)*, Oxford, UK, September 2003.

DeRose, S., Maler, E., and Daniel, R. eds. 2001. XML Pointer Language (XPointer) Version 1.0. W3C candidate recommendation.

Hahn, U., Romacker, M., and Schulz, S. 2002. Creating Knowledge Repositories from Biomedical Reports: The MEDSYNDIKATE Text Mining System. In *Proceedings of the Pacific Symposium on Biocomputting 2002*, pp. 338-349.

Hirschman, L., and Chinchor, N. 1997. MUC-7 Coreference Task Definition, Version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)* 1997.

Kawazoe, A. and Collier, N. 2003. An Ontologically-motivated Annotation Scheme for Coreference, in *proceedings of the International Workshop on Semantic Web Foundations and Application Technologies*, Nara, Japan, 11th March, 2003.

Lassila, O., and Swick, R. eds. 1999. Resource Description Framework (RDF) Model and Syntax Specification. Recommendation, W3C, Feb. 1999.

Liu, Y., Lazar, N., Rothfus, W.E., Dellaert, F., Moore, A., Schneider, J., and Kanade, T. 2004. Semantic based Biomedical Image Indexing and Retrieval. *Trends and Advances in Content-Based Image and Video Retrieval*, Shapiro, Kriegel, and Veltkamp, ed., February, 2004.

Noy, N. F., Fergerson, R. W. and M. A. Musen, M. A. 2000. The knowledge model of Protégé-2000: combining interoperability and flexibility. In *Proceedings of the 2nd International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000)*, Juan-les-Pins, France, pages 1–20, 2000.

Shotton, D. M. 2003. The BioImage Database: Multidimensional research images and their relationship to the wider online research information environment for the life sciences (best viewed with Internet Explorer). *Proc. E-BioSci / ORIEL Open Meeting - Biological Information Management: Challenges and Choices.* National e-Science Centre, Edinburgh, April, 2003.

van Deemter, K. 1998. Representations for Multimedia coreference". In *Proc. of Workshop on Combining ai and Graphics for the Interface of the Future*, held in conjunction with ecai'98, Brighton, uk, pp.1-9.

Yamada, K., and Nakagawa, H. 2002 Identification of Coreference between Names and Faces. *IPSJ Journal Vol.* 43 (6), pp: 1890-1898. Information Processing Society of Japan.

[1] http://research.nii.ac.jp/~collier/papers/KES2003.pdf