An Ontologically-motivated Annotation Scheme for Coreference

Ai Kawazoe*[†] and Nigel Collier*
*National Institute of Informatics (NII)
National Center of Sciences
2-1-2 Hitotsubashi Chiyoda-ku,
Tokyo 101-8430, Japan
[†]Kyushu University
6-10-1 Hakozaki, Higashi-ku
Fukuoka 812-8581, Japan
{zoeai, collier}@nii.ac.jp

Abstract

This paper provides an overview of the annotation scheme for coreference which is now being developed with the aim of applying it to the Semantic Web. We will state that our scheme assumes the theoretically-motivated view of coreference which we call the "symmetric view", and that the scheme enables to link annotations directly to the ontology and allows annotators from various domains to make consistent annotations.

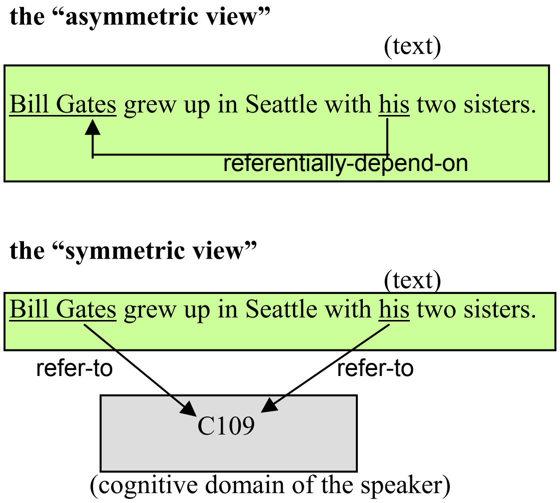
1. Introduction

In this paper we present an annotation scheme for coreference which is ontologically-motivated, and based on the result of recent theoretical studies. The work allows both linguists as well as non-linguists to annotate coreference relations between expressions in a text that are considered as instances of classes in the Semantic Web ontology.

Although in the literature many coreference annotation schemes have been proposed for many purposes, such as MUC-7 coreference annotation (Message Understanding Conference; Hirschman and Chinchor 1997), MATE annotation scheme for coreference (Poesio 2000, Poesio, Brunesseaux and Romary 1999, Davies et al. 1998), MEDSTRACT anaphoric annotation (Castaño, Zhang, and Pustejovsky 2002, Pustejovsky et al. 2002), and UCREL anaphoric annotation (Figelstone 1992), surprisingly few schemes have been applied to general annotation and none specifically to the Semantic Web. We consider that this is because most of existing schemes annotate coreference as a dependency of anaphoric expressions to their antecedents, and thus a purely document-internal relation. In this paper, we take a theoretically motivated view that coreference is basically a relation between a referred object and the expressions each of which independently refers to the object, and based on this view we propose an out-of-text annotation of coreference which allows direct linkage to the public ontology.

2. "Asymmetric" vs. "Symmetric" views of annotation

In the literature of theoretical linguistics, there are two different views for the phenomena of coreference. One of them is the view that coreference is a collection of text-internal relation between two expressions such that one of them is "referentially dependent" on the other, what we call here as the "asymmetric view." The other one is the view that coreference is a phenomenon that more than two expressions independently refer to the same object in the world or in some cognitive domain: let us call it the "symmetric view." With this view, coreference is not a relation among linguistic expressions, but an instance of the relation between the expressions and their referents in the outside of the text.


the "asymmetric view"

the "symmetric view"

(text)
Bill Gates grew up in Seattle with his two sisters.
referentially-depend-on
(text)
Bill Gates grew up in Seattle with his two sisters.
refer-to
refer-to
C109
(cognitive domain of the speaker)

Figure 1.
The "asymmetric" and "symmetric" views of coreference

In the recent studies such as (Ueyama 1998) and (Hoji et al. 2000), the distinction between coreference and typical dependency anaphora such as bound variable anaphora has been made clear, and it is claimed that coreference phenomena should be analyzed in terms of the "symmetric" view.

Most of the existing coreference annotation schemes annotate coreference as an asymmetric relation among coreference occurrences. For example, in MUC-7, MATE and MEDSTRACT, a pair of coreference occurrences is annotated as having an intra-document relation that one of them is an antecedent and the other one a dependent term.

However, it cannot be denied that this "asymmetric view" would cause the difficulty in applying these schemes to the use on the Semantic Web. First of all, it is not at all easy for annotators to decide which expression is the antecedent of which one. MUC-7 and MATE allows annotators to choose antecedents freely. In the annotation sample of MUC-7, the nearest coreferential expression is simply chosen as an antecedent. However, this sometimes leads to counterintuitive results: for example, it would allow a pronoun to be the antecedent of a name. In MEDSTRACT such unfavorable situations seem to be avoided by setting up a hierarchy among referential expressions and constraints on the choice of antecedents, but maintaining the memory of the hierarchy and the constraints are not an easy task.

Secondly, it is also difficult for "asymmetric" schemes to create coreference chains between annotations in different documents. When creating such chains with "asymmetric" annotations, it is necessary for annotators to decide which one of coreference occurrences becomes a "conventional form." For example, they need to choose the most appropriate label from "Gates," "Bill Gates," and "William H. Gates," but it may cause confusions and inconsistencies among annotators.

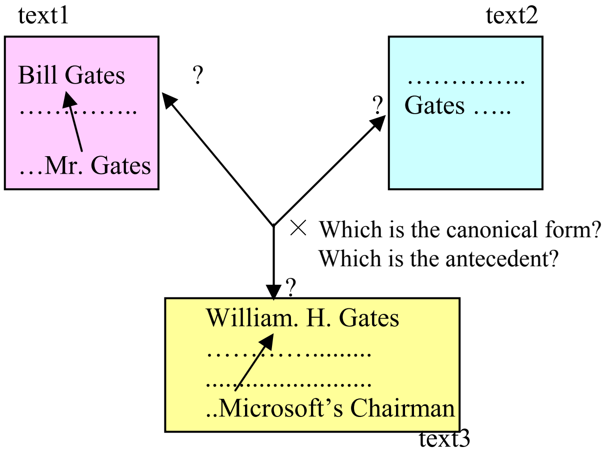


Figure 2.
Asymmetric annotation for inter-text coreference

Regarding these difficulties in the "asymmetric" annotations, it seems reasonable to make a shift to the "symmetric" view, which is motivated by the theoretical studies. With this view, choosing antecedents and conventional forms among coreference occurrences is unnecessary. Further, most importantly, "symmetric" annotation scheme makes it possible for coreference relations to be linked to the public ontology. With this linkage, a semantic consistency is achieved when a computer tries to understand and reason with the entities in a text.

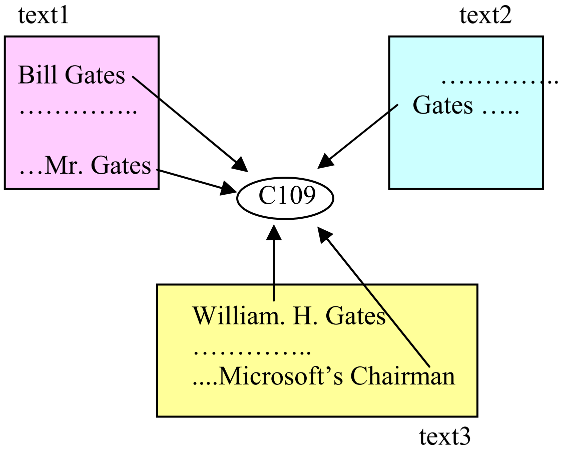


Figure 3.
Symmetric annotation for inter-text coreference
(Cf. Figure 2)

3. The scheme based on the "symmetric view"

Here we describe our annotation scheme for coreference. Our basic approach is as follows:

- take all coreference occurrences as equal in that they each independently refer to a concept.
- consider annotations as the first class objects and allow them to occur independent of the base document.

Thus we have a coreference annotation document separated from the original text along the same lines as (Collier et al. 2002), and in the document we can describe the relations among referred concepts, e.g. a membership relation, etc.

A selected view of an annotation can be seen in Figure 4. As shown in the figure, the annotation property *context* relates the annotations to the resource (a Web page) and takes on an XPointer value (DeRose, Maler and Daniel 2000). The property *identity_id* relates each of the annotations to a concept in the annotation document. Since we provide concepts with IDs such as "C1024," and specify them by the set of expressions which refer to them, we can prevent the potential confusion and inconsistencies on deciding which surface form should be the “conventional form” of the concept.

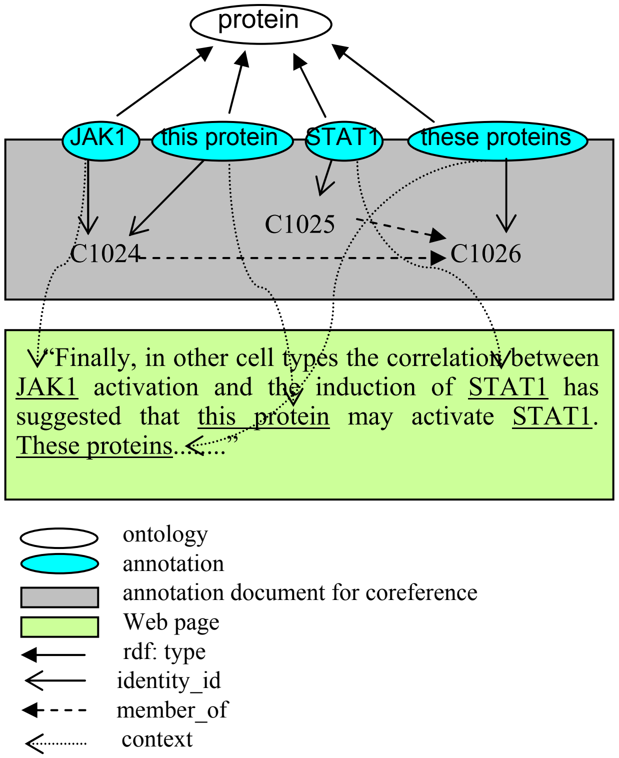

protein
JAK1
this protein
STAT1
these proteins
C1024
C1025
C1026
“Finally, in other cell types the correlation between JAK1 activation and the induction of STAT1 has suggested that this protein may activate STAT1. These proteins......”
ontology
annotation
annotation document for coreference
Web page
rdf: type
identity_id
member_of
context

Figure 4. Overview of the annotation

Further, we can link each coreference annotation to the public ontology in terms of the relation defined in RDF (Lassila and Swick 1999). As shown in Figure 4, the implementation of coreference relations between *JAK1* and *this protein* helps to maximize information about the object in the text. This suggests the possibility that our scheme can make contributions to ontology modification and construction.

Our most basic aim within PIA project (Collier and Takeuchi 2002, Collier, Takeuchi and Tsuji 2001) is to produce annotated texts that are useful for machine learning of coreference resolution. In order to maintain consistency of annotation for this aim, at this first stage we annotate only “identity of reference” relations among Noun Phrases (NPs), thus we do not annotate any instances of bound variable anaphora, E-type anaphora, zero anaphora, bridging references, and other coreference cases which involves events denoted by verb phrases or propositions denoted by sentences. In addition, we subtype the coreference occurrences (such as *name, alias, pronoun, definite* and *indefinite*, etc.) with the annotation class *type*, in order to enable machines to learn coreference resolution in a step-by-step manner; i.e., the learning begins with the easiest subtypes of coreference and move gradually to more difficult ones. We are conducting a survey on the easiness/ difficulty for each subtypes of coreference, and planning to develop our scheme according to the result of the survey.

4. A comparison with another “symmetric” scheme

The scheme we present here is not the first 'symmetric' annotation scheme for coreference in the literature, e.g. parts of MMAX (Multi-Modal Annotation in XML; Müller and Strube 2001a, 2001b). This scheme divides the coreference annotation process into the following two steps:

1. annotate an anaphoric expression that co-specifies (corefers) with all other markables already in the set of co-specifying expressions. (obligatory)
2. specify the anaphoric expression's exact antecedent. (optional)

Our scheme is slightly different to MMAX in that we do not adopt their second step, i.e. the optional specification of antecedents. We do not view the definition of 'antecedent' as being simple or indeed an essential step for our coreference requirements at this time. Furthermore, when information about the nature of the antecedent is necessary for coreference resolution the definition of the type of antecedent will need to change influencing the resolution algorithm that will be applied. Thus, we do not specify antecedents in our annotation scheme.

5. Conclusion

In this study we have described our annotation scheme based on the view supported by theoretical studies, and discussed its usefulness on the Semantic Web.

We are constructing annotation guidelines and a test set of annotated EMBO Journal articles in the molecular biology domain. We are also planning to provide software support for annotators within Open Ontology Forge which is now being produced. In the future work we will extend our scheme to other languages such as Japanese.

References

Castaño, R., Zhang, J., and Pustejovsky, J. 2002. Anaphora Resolution in Biomedical Literature. International Symposium on Reference Resolution for Natural Language Processing, Alicante, Spain.
(URL: http://medstract.org/papers/coreference.pdf)

Collier, N., Takeuchi, K., Nobata, C., Fukumoto, J., and Ogata, N. 2002. Progress on Multi-lingual Named Entity Annotation Guidelines using RDF(S). In *Proceedings of the Third International Conference in Language Resources and Evaluation*, 2074-2081. Las Palmas de Gran Canaria, Spain.

Collier, N., and Takeuchi, K. 2002. PIA-Core: Semantic Annotation through Example-based Learning. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, 1611-1614. Las Palmas, Spain.

Collier, N., Takeuchi, K., and Tsuji, K. 2001. The PIA Project: Learning to Semantically Annotate Texts from an Ontology and XML-Instance Data. In *Position Paper Proceedings of the First Semantic Web Working Symposium (SWWS'2001)*, 8-9. Stanford University, California, USA.

Davies, S., Poesio, M., Bruneseaux, F., and Romary, L. 1998. Annotating Coreference in Dialogues: Proposal for a Scheme for MATE, First draft. (URL:http://www.hcrc.ed.ac.uk/~poesio/MATE/anno_manual.html)

DeRose, S., Maler, E., and Daniel, R. eds. 2001. XML Pointer Language (XPointer) Version 1.0. W3C candidate recommendation, 11th September 2001. (URL: http://www.w3.org/TR/2001/CR-xptr-20010911/)

Fligelstone, S. 1992. Developing a Scheme for Annotating Text to Show Anaphoric Relations. In: Leitner, G. (ed.) *New Directions in Corpus Linguistics*: 153-170. Berlin: Mouton de Gruyter.

Hirschman, L., and Chinchor, N. 1997. MUC-7 Coreference Task Definition, Version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7).*(URL:http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html)

Hoji, H., Kinsui, S., Takubo, Y., and Ueyama, A. 2000. Demonstratives, Variables, and Reconstruction Effects. In *Proceedings of the Nanzan GLOW (The Second GLOW Meeting in Asia)*, 141-158.

Lassila, O., and Swick, R. eds. 1999. Resource Description Framework (RDF) Model and Syntax Specification. Recommendation, W3C, Feb. 1999. (URL: http://www.w3.org/TR/1999/REC-rdf-syntax)

Müller, C., and Strube, M. 2001a. Annotating Anaphoric and Bridging Expressions with MMAX. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, 90-95. Aalborg, Denmark.

Müller, C., and Strube, M. 2001b. MMAX: A Tool for the Annotation of Multi-modal Corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 45-50. Seattle, Wash.

Poesio, M. 2000. Coreference. In Mengel, A., Dybkjaer, L., Garrido, J.M., Heid, U., Klein, M., Pirrelli, V., Poesio, M., Quazza, S., Schiffrin, A., and Soria, C. *MATE Dialogue Annotation Guidelines*. (URL: http://www.ims.uni-stuttgart.de/projekte/mate/mdag/cr/cr_1.html)

Poesio, M., Bruneseaux, F., and Romary, L. 1999. The MATE Meta-scheme for Coreference in Dialogues in Multiple Language. In *Proceedings of the ACL Workshop on Standards for Discourse Tagging*, 65-74.

Pustejovsky, J., Castaño, J., Saurí, R., Rumshisky, A., Zhang, J., Luo, W. 2002. Medstract: Creating Large-scale Information Servers for Biomedical Libraries. In *ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain*. Philadelphia, PA. (URL: http://www.medstract.org/papers/acl2002-4.pdf)

Ueyama, A. 1998. Two Types of Dependency. Doctoral dissertation, University of Southern California, distributed by GSIL publications, USC, Los Angeles.