

Annotation of all coreference in biomedical text: Guideline selection and adaptation

K. Bretonnel Cohen^{1,2}, Arrick Lanfranchi², William Corvey²,
William A. Baumgartner Jr.¹, Christophe Roeder¹, Philip V. Ogren^{1,3},
Martha Palmer², Lawrence E. Hunter¹

1: Center for Computational Pharmacology
University of Colorado School of Medicine
Aurora, Colorado, USA

2: Department of Linguistics
University of Colorado at Boulder
Boulder, Colorado, USA

3: Department of Computer Science
University of Colorado at Boulder
Boulder, Colorado, USA

Abstract

This paper describes an effort to build a corpus of full-text journal articles in which every co-referring noun phrase is annotated. The identity and appositive relations were marked up. Several annotation schemas were evaluated and are described here; the OntoNotes guidelines were selected. Biomedical journal articles required a number of adaptations to the OntoNotes guidelines—mainly doing away with the notion of generics, which also had implications for the handling of nominal modifiers. Domain experts and linguists were evaluated with respect to their ability to function as annotators, and both were found to be effective. Progress is reported with about one third of the project done; inter-annotator agreement at this stage is 0.684 by the MUC metric.

1. Introduction

The Colorado Richly Annotated Full Text (CRAFT) corpus is a set of 97 full-text journal articles that is currently being annotated in a joint project between the University of Colorado School of Medicine and the Linguistics Department of the University of Colorado at Boulder. The corpus contains about 597,000 words of text and is being annotated for a number of types of linguistic information, including part of speech and treebanking, and semantic information, including a variety of types of named entities. Additionally, we have included some discourse structure annotation in the form of coreference.

This project is unlike any other that we are aware of in that it includes marking *all* coreferential relations of identity and apposition between *all* noun phrases in the full text of a large body of publicly available biomedical publications. Other work (Gasperin, 2006; Gasperin et al., 2007) has done annotation of full biomedical text, but only of noun phrases referring to biological entities. In contrast, we mark up coreference between any and all noun groups in the documents. IIR has done full coreference and appositive annotation of full-text journal articles, but on a smaller set, and has not made them publicly available. The CRAFT corpus is about twice the size of the IIR document set and will be made publicly available. This is the most ambitious project of its kind of which we are aware.

2. Methods and Results

2.1. Selection of annotation guidelines

One of the desiderata of the project was to contribute to the development of a standard for coreference annotation by adopting a pre-existing set of annotation guidelines, rather

than developing our own de novo. To this end, a number of publicly available annotation guidelines were evaluated. (We did not consider projects that only tackled pronominal anaphora, to the exclusion of full noun phrase coreference, or projects that tackled clinical data.) We were initially relatively agnostic as to desiderata, other than that we wanted the guidelines to include a full range of coreferential and bridging anaphoric relations, as well as part/whole relationships.

2.1.1. OntoNotes

OntoNotes (Hovy et al., 2006) is a large, multi-center project to create a multi-lingual, multi-genre corpus annotated at a variety of linguistic levels, including coreference (Pradhan et al., 2007). As part of the OntoNotes project, the BBN Corporation prepared a set of annotation guidelines. They are not publicly available. (The version currently available online at the Linguistic Data Consortium is a full numbered version out of date, compared to the version that we used.) However, (Pradhan et al., 2007) gives the flavor of the approach.

The OntoNotes guidelines include events, pronominal and full anaphora and coreferents, and verbs as markables. Predicative nouns are not treated as coreferential. There is a separate relation for appositives. Nominal premodifiers are markables, with some restrictions that we discuss below in the section on domain-specific changes to the guidelines. The guidelines only include the identity and appositive relations, with set membership and part/whole not included, but this turned out not to be relevant to our project, since funding constraints precluded annotating these relations.

2.1.2. Gasperin

Gasperin (Gasperin, 2006; Gasperin et al., 2007) is the only previous attempt that we are aware of to annotate the full text of biomedical journal articles. Her project involved annotating five such articles. The annotation guidelines reflect a nuanced attempt to reflect the semantics of the biological domain. They do this in part by only selecting biomedical entities as markables, but more importantly by defining a domain-relevant set of relations. In addition to coreference, the relations are three types of associative relations: homology; related “biotype” (e.g. a gene and its protein or a gene and a subsequence of that gene); and the set/member relation.

Appositives and predicative nouns are both treated as coreferential. Premodifiers are not specifically addressed in the guidelines.

Predicative nouns are treated as coreferential. Pronouns are excluded completely. Probably the most striking aspect of Gasperin’s guidelines is its definition of markables; they are limited to “bio-entities.” This was a major mismatch with our goals, which included annotating all coreferential entities.

2.1.3. GENIA project at IIR

The Institute for Infocomm Research in Singapore and the Department of Computer Science spearheaded an annotation project involving the 2,000 abstracts in the GENIA corpus (Yang et al., 2004a; Yang et al., 2004b) as well as 43 full papers. They annotated four relations: identity, appositives, pronouns (considered separately from other identity relations), and relative pronouns. The annotation of markables included a minimal string that would suffice for identification. As in the case of the other projects besides OntoNotes, they annotated only nouns, not verbs; the guidelines allowed for these, but they did not find any while preparing the guidelines. Premodifiers were not considered markables. Predicative nouns were marked when they were definite.

2.1.4. MUC7

The groundbreaking MUC7 guidelines (Hirschman, 1997) have underlaid a number of subsequent coreference annotation projects. The scheme covered only a single relation, identity. The annotation of markables included a minimal string that would suffice for identification of the corefering element. Noun phrases and pronouns were included as markables. Gerunds were excluded as markables. Appositives and predicate nominals were both marked, as coreferential. Prenominal modifiers were annotated only in the case where they could be linked to something other than another prenominal modifier.

A number of authors have critiqued various aspects of the MUC7 coreference annotation guidelines, e.g. (van Deemter and Kibble, 2001). A strong point of the MUC7 guidelines is that they make an attempt to deal with changing numerical values, as in *The results of this analysis showed that the statistical support for the linkage of the C57BL/6 locus on Chromosome 3 for ANA increased from logarithm of odds (LOD) 5.4 to LOD 6.4*, which other guidelines do not, although even the MUC7 organizers were not happy with their approach to this. We note that

a number of other problems pointed out with the MUC7 guidelines, including situations where the referential status of the markable is unclear, are present in all of the other guidelines that we examined as well.

2.2. Domain-specific changes to the guidelines

After reviewing the pre-existing guidelines, senior annotators marked up a sample full-text article, following the OntoNotes guidelines. We found the OntoNotes guidelines to be a good match to our conception of how coreference should be annotated, and noted that they have responded to a number of critiques of earlier guidelines. For example, compared to the MUC-7 guidelines, the treatment of appositives in terms of heads and attributes rather than separate mentions is an improvement in terms of referential status, as is the handling of predicative nouns. The inclusion of verbs and events is a desirable increase in scope. The guidelines are more detailed, as well. They were also attractive from a political point of view—we wanted to adopt a widely accepted set of guidelines, and they seemed like a good candidate for this due to their association with a large project. We adopted them; however, the nature of the biomedical domain required a major adaptation of the guidelines.

2.2.1. Generics

The OntoNotes guidelines make crucial reference to a category of nominal that they refer to as a *generic*. Generics include:

- bare plurals
- indefinite noun phrases
- abstract and underspecified nouns

The status of generics in the annotation guidelines is that they cannot be linked to each other via the IDENTITY relation. They can be linked with subsequent non-generics, but never to each other, so every generic starts a new IDENTITY chain (assuming that it does corefer with subsequent markables).

The notion of a generic is problematic in the biomedical domain. The reason for this is that any referring expression in a biomedical text is or should be a member of some biomedical ontology, be it in the set of Open Biomedical Ontologies, the Unified Medical Language System, or a nascent ontology. As such, it has the status of a named entity. To take an example from BBN, consider the status of *cataract surgery* in the following:

Allergan Inc. said it received approval to sell the PhacoFlex intraocular lens, the first foldable silicone lens available for **cataract surgery**. The lens’ foldability enables it to be inserted in smaller incisions than are now possible for **cataract surgery**.

According to the OntoNotes guidelines, *cataract surgery* is a generic, by virtue of being abstract or underspecified, and therefore the two noun phrases are not linked to each other via the IDENTITY relation. However, *cataract surgery*

is a concept within the Unified Medical Language System (Concept Unique Identifier C1705869), where it occurs as part of the SNOMED Clinical Terms. As such, it is a named entity like any other biomedical ontology concept, and should not be considered generic. Indeed, it is easy to find examples of sentences in the biomedical literature in which we would want to extract information about the term *cataract surgery* when it occurs in contexts in which the OntoNotes guidelines would consider it generic:

- *Intravitreal administration of 1.25 mg bevacizumab at the time of cataract surgery was safe and effective in preventing the progression of DR and diabetic maculopathy in patients with cataract and DR.* (PMID 19101420)
- *Acute Endophthalmitis After Cataract Surgery: 250 Consecutive Cases Treated at a Tertiary Referral Center in the Netherlands.* (PMID 20053391)
- *TRO can present shortly after cataract surgery and lead to serious vision threatening complications.* (TRO is thyroid-related orbitopathy; PMID 19929665).

In these examples, we might want to extract an IS_ASSOCIATED_WITH relation between <bevacizumab, cataract surgery>, <acute endophthalmitis, cataract surgery>, and <thyroid-related orbitopathy, cataract surgery>. This makes it important to be able to resolve coreference with them.

Thus, our project’s guidelines do not consider there to be generics as such in the genre and domain that it is concerned with¹.

2.2.2. Prenominal modifiers

A related issue concerned the annotation of prenominal modifiers. The OntoNotes guidelines call for prenominal modifiers to be annotated only when they are proper nouns. However, since we considered all entities to be named entities, our guidelines called for annotation of prenominal modifiers regardless of whether or not they were proper nouns, as such.

2.3. The annotation schema

2.3.1. Noun groups

The basic unit of annotation in the project is the base noun phrase. We defined this as one or more nouns and any sequence of leftward determiners, adjectives, and conjunctions not separated by a preposition or other noun-phrase-delimiting part of speech; and rightward modifiers such as relative clauses and prepositional phrases.

Thus, all of the following would be considered base noun phrases:

- *striatal volume*
- *neural number*

- *striatal volume and neural number*
- *the structure of the basal ganglia*
- *It*

These were not pre-annotated—the annotators selected their spans themselves. This is one potential source of lack of interannotator agreement. Base noun phrases were annotated only when they participated in one of the two relationships that we targeted.

2.3.2. Definitions of the two relations

The two relations that are annotated in the corpus are the IDENTITY relation and the APPOSITIVE relation. The identity relation holds when two units of annotation refer to the same thing in the world. The appositive annotation holds when two noun phrases or a noun phrase and an appositive are adjacent and not linked by a copula or other linking word.

2.3.3. Details of the annotation schema

More specifically, the annotation schema is defined as:

IDENTITY chain An IDENTITY chain is a set of base noun phrases and/or appositives that refer to the same thing in the world. It can contain any number of elements.

Base noun phrase Discussed above.

APPOSITIVE relation An appositive instance has two elements, a head and a set of attributes. The set of attributes may contain just a single element (the prototypical case). Either the head or the attributes may themselves be appositives.

Nonreferential pronoun All nonreferential pronouns are included in this single class.

Thus, an example set of annotations would be:

All brains analyzed in this study are part of [the Mouse Brain Library]_a ([MBL]_b). [The MBL]_c is both a physical and Internet resource. (PMID 11319941)

- APPOSITIVE chain: *The Mouse Brain Library_a, MBL_b*
- IDENTITY chain: *Mouse Brain Library_a, The MBL_c*

2.4. Training of the annotators

We hired and trained biologists and linguists as a group. Annotators were given a lecture on the phenomenon of coreference and on how to recognize coreferential and appositive relations, as well as nonreferential pronouns. They were then given a non-domain-specific practice document. Following a separate session on the use of the annotation tool, they were given an actual document to annotate. This document is quite challenging, and exercised all of the necessary annotation skills. We began with paired annotation, then introduced a second document for each annotator to mark up individually. Once annotators moved on to individual training annotation, they met extensively with a senior annotator to discuss questions and review their final annotations.

During the initial training phase, we paired biologists with linguists and had them work on the same article independently, then compare results. This turned out to be an unnecessary step, and we soon switched to having annotators work independently from the beginning.

¹Note that in our guidelines, as in the OntoNotes project, indefinite noun phrases are used to start new IDENTITY chains, and are not linked with previous markables, but this is because they are discourse-new, not because we consider them to be generics.

2.5. Two populations of annotators

We hired two very different types of annotators—linguistics graduate students, and biologists at varying levels of education and with varying specialties. Impressionistically, we did not notice any difference in their performance. The biologists were able to grasp the concept of coreference, and the linguists did not find their lack of domain knowledge to be an obstacle to annotation. Both the biologists and the linguists had opportunities to clarify issues via email and meetings with the senior annotators.

2.6. The annotation process

Most articles are single-annotated, but a subset of fifteen will be double-annotated by random pairs of annotators to calculate inter-annotator agreement.

The length of the articles means that a single IDENTITY chain can extend over an exceptionally long distance. To cope with this, annotators typically marked up single paragraphs as a whole, and then linked entities in that paragraph to earlier mentions in the document.

In the case of questions, annotators had access to senior annotators via email and meetings.

Annotation was done using Knowtator, a Protégé plug-in (Ogren, 2006a; Ogren, 2006b).

2.7. Progress to date

Table 1 shows the total number of documents, IDENTITY chains, APPOSITIVE chains, nonreferential noun phrases, and base noun phrases with about one third of the project done. (In calculating these numbers, when a document had been double-annotated, we took the average of the two annotators for that document.) These numbers give some idea of the scale of the task of annotating all coreferential noun phrases in full-text scientific journal articles, along with the data in Table 2, where we see the averages. There we note especially that the average time to annotate a single article is twenty hours. This number should be useful for estimating the funding needed for future annotation projects of this sort.

The low number of nonreferential pronouns is striking, but accords with the work of Gasperin, who reported numbers of nonreferential pronouns in her full-text articles that were so low that she omitted them from the annotation schema.

In contrast, the high number of base noun phrases is quite striking. Recall that our process includes the annotators marking the boundaries of the base noun phrases themselves; this high number suggests that a significant time savings might be realized by pre-marking the syntactic constituents of the papers. We also note from early work by Hirschman et al. that this would likely increase our inter-annotator agreement.

Average inter-annotator agreement over a set of ten articles is .684 by the MUC metric. We give a number of other metrics in Table 3 (MUC, (Vilain et al., 1995), B3, (Bagga and Baldwin, 1998), CEAF, (Luo, 2005), and Krippendorff’s alpha (Passonneau, 2004; Krippendorff, 1980)). We note that the value for Krippendorff’s alpha is lower than the 0.67 that Krippendorff indicates must be obtained before values can be conclusive, but no other IAA values

Documents annotated	31
Total IDENT chains	6,650
Total APPOS chains	1,230
Total nonreferential pronouns	296
Total base noun phrases	27,870

Table 1: **Total documents and markables to date.** Base noun phrase count includes only those base noun phrases included within an IDENT or APPOS chain.

Average time per document	20 hours
Average IDENT chains per document	215
Average APPOS chains per document	40
Average nonreferential pronouns per document	10
Average base noun phrases per document	899

Table 2: **Average markables to date.** Base noun phrase count includes only those base noun phrases included within an IDENT or APPOS chain.

for projects using the OntoNotes guidelines have been published to compare these numbers to. Note again that these are preliminary numbers representing progress to date only and do not represent the inter-annotator agreement values for the completed project.

3. Discussion

We found that published and unpublished annotation guidelines for coreference differ widely with respect to the definitions of markables, handling of appositives, handling of predicative noun phrases, handling of nominal premodifiers, and the inclusion of other relations in addition to coreference. A number of them are compared and contrasted here.

We also found that a candidate for a standard set of coreference guidelines, the OntoNotes guidelines, required considerable adaptation to fit the nature of the biomedical domain, particularly with respect to the notion of generics, which play a large role in the OntoNotes guidelines. The description of our annotation process describes how these guidelines can be incorporated into an active annotation project, and our report on progress to date gives an early indication of the scale of the data that will result from annotation to these guidelines.

Acknowledgements

We thank the BBN Corporation for sharing the OntoNotes coreference annotation guidelines with us, and the other

Metric	Average
MUC	.684
Class-B3	.858
Entity-B3	.750
Mention-based CEAF	.644
Entity-based CEAF	.480
Krippendorff’s alpha	.619

Table 3: **Inter-annotator agreement values for a sample of ten documents.** A variety of metrics is reported.

writers of guidelines for making their guidelines publicly available. We also thank Guergana Savova and Jiaping Zheng of the Mayo Clinic for sharing their inter-annotator agreement code with us. Two anonymous reviewers improved the paper.

4. References

- Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LREC '98)*, pages 563–566.
- Caroline Gasperin, Nikiforos Karamanis, and Ruth Seal. 2007. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proceedings of DAARC 2007*.
- Caroline Gasperin. 2006. Semi-supervised anaphora resolution in biomedical texts. In *Linking natural language processing and biology: towards deeper biological literature analysis*, pages 96–103. Association for Computational Linguistics.
- L. Hirschman. 1997. MUC-7 coreference task definition.
- Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Human Language Technology Conference of the NAACL Companion Volume*, pages 57–60.
- Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology (Commtext Series)*. SAGE Publications, September.
- Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 25–32.
- Philip Ogren. 2006a. Knowtator: a Protege plugin for annotated corpus construction. In *HLT-NAACL 2006 Companion Volume*.
- Philip Ogren. 2006b. Knowtator: a plug-in for creating training and evaluation data sets for biomedical natural language systems. In *The International Protege conference*, pages 73–76.
- Rebecca J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of the Language Resources and Evaluation Conference*.
- Sameer S. Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: identifying entities and events in OntoNotes. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, pages 446–453.
- Kees van Deemter and Rodger Kibble. 2001. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637.
- Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 45–52.
- Xiao Feng Yang, Jian Su, Guo Dong Zhou, and Chew Lim Tan. 2004a. A NP-cluster based approach to coreference resolution. In *Proceedings of 20th International Conference on Computational Linguistics (COLING 2004)*, pages 226–232.
- Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2004b. Improving noun phrase coreference resolution by matching strings. In *IJCNLP04*, pages 326–333.