Building a Coreference-Annotated Corpus from the Domain of Biochemistry

Riza Theresa Batista-Navarro[1,2,3,†] and **Sophia Ananiadou**[1,2,††]
[1]National Centre for Text Mining, University of Manchester, United Kingdom
[2]School of Computer Science, University of Manchester, United Kingdom
[3]Department of Computer Science, University of the Philippines Diliman, Philippines
†batistar@cs.man.ac.uk, ††sophia.ananiadou@manchester.ac.uk

Abstract

One of the reasons for which the resolution of coreferences has remained a challenging information extraction task, especially in the biomedical domain, is the lack of training data in the form of annotated corpora. In order to address this issue, we developed the *HANAPIN* corpus. It consists of full-text articles from biochemistry literature, covering entities of several semantic types: chemical compounds, drug targets (e.g., proteins, enzymes, cell lines, pathogens), diseases, organisms and drug effects. All of the co-referring expressions pertaining to these semantic types were annotated based on the annotation scheme that we developed. We observed four general types of coreferences in the corpus: sortal, pronominal, abbreviation and numerical. Using the MASI distance metric, we obtained 84% in computing the inter-annotator agreement in terms of Krippendorff's alpha. Consisting of 20 full-text, open-access articles, the corpus will enable other researchers to use it as a resource for their own coreference resolution methodologies.

1 Introduction

Coreferences are linguistic expressions referring to the same real-world entity (Jurafsky and Martin, 2009). The process of grouping all co-referring expressions in text into respective coreference chains is known as *coreference resolution*. It was introduced as one of the tasks of the sixth Message Understanding Conference (MUC-6) in 1995 (Grishman and Sundheim, 1995) and is one of the information extraction tasks which have remained a challenge to this day. One of the reasons it is still considered an unresolved problem especially in the biomedical domain is the lack of coreference-annotated corpora which are needed for developing coreference resolution systems.

There exist only a handful of biomedical corpora which are annotated with coreference information. We have conducted a review of each of them, taking into consideration their sizes, document composition, domain, types of markable entities, types of coreference annotated, availability, and reliability in terms of inter-annotator agreement. Of these, only two corpora have been used in coreference resolution systems developed outside the research group that annotated them: MEDSTRACT (Castano et al., 2002), and the MEDCo[1] corpus of abstracts which was used by the different teams who participated in the Coreference Supporting Task of the BioNLP 2011 Shared Task[2]. These two corpora are widely used, despite the fact that they are composed only of abstracts.

Previous studies have shown the advantages of utilising full-text articles rather than abstracts in information extraction systems (Shah et al., 2003; Schumie et al., 2004; Cohen et al., 2010a). Furthermore, recent research on fact extraction (McIntosh and Curran, 2009) has demonstrated the need for processing full-text articles when identifying coreferent expressions pertaining to biomedical entities.

[1]http://nlp.i2r.a-star.edu.sg/medco.html
[2]http://sites.google.com/site/bionlpst/home/protein-gene-coreference-task

Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pages 83–91,
Portland, Oregon, USA, June 23-24, 2011.

However, coreference-annotated corpora composed of full-text articles are not readily accessible. Currently, only the FlySlip corpus (Gasperin et al., 2007) is available for download. In this corpus, only gene-related entities were considered for coreference annotation. Thus, there is a need for developing full-text corpora with coreference annotations for more semantic types. This is currently being addressed by the CRAFT project (Cohen et al., 2010b) which seeks to develop a corpus of full-text articles with coreference annotations for more types of entities; it was not explicitly stated, however, exactly which types are being covered. Similarly, we are developing a corpus of full-text articles with coreference annotations, but to further the aim of covering as many semantic types as possible, we selected a domain that covers a variety of semantic concepts. Research literature from this biochemistry subdomain, *marine natural products chemistry*, contains references pertaining to chemical compounds, organisms, drug targets such as proteins, enzymes, nucleic acids, tissues, cells, cell components, cell lines and pathogens, drug effects, as well as diseases. We cover a number of entity types with the intention of providing more insight into how to disambiguate coreferring expressions of different semantic types.

An annotation scheme was developed, taking into consideration the coreference types which have been observed from the corpus, namely: sortal, pronominal, numerical and abbreviation. Three chemistry graduates were employed to annotate the corpus. To determine the reliability of the resulting annotations, we measured inter-annotator agreement in terms of Krippendorff's alpha.

2 Related Work

Coreference is often associated with the phenomenon of *anaphora* which is characterised by an expression (called an *anaphor*) that points back to an entity previously mentioned in the same discourse (called *antecedent*). Anaphora resolution is the process of determining the antecedent of an anaphor. While the output of anaphora resolution is a set of anaphor-antecedent pairs, that of coreference resolution is a set of coreference chains which can be treated as equivalence classes. Despite this difference, an overlap between them may be observed in several cases. Often, a number of anaphor-antecedent pairs from a discourse are coreferential or refer to the same entity in the same domain, and may be placed in the same coreference chain. For this reason, we also included in our review of biomedical corpora those which were annotated with anaphora information and refer to them henceforth as coreference-annotated corpora.

We determined the types of coreference annotated in each corpus we have reviewed, adapting Mitkov's classification of anaphora (Mitkov et al., 2000) which is also applicable to coreference. *Nominal coreference* is characterised by co-referring expressions pertaining to a noun. It is further divided into *pronominal coreference* and *sortal coreference* which use a pronoun and a lexical noun phrase, respectively, as co-referring expressions. Unlike nominal coreference, *verbal coreference* is characterised by co-referring expressions pertaining to verbs. Both nominal and verbal coreference can be broadly categorised according to the kind of relation as *direct* or *indirect*. In direct coreference, co-referring expressions are related by identity, synonymy or specialisation; in indirect coreference, they are related by associative relations such as meronymy or holonymy for nouns, and troponymy or entailment for verbs. Annotation of indirect coreference is usually more challenging as it requires more specialised domain knowledge.

Presently, there are five (5) different biomedical corpora which are annotated with coreference information: MEDSTRACT (Castano et al., 2002), MEDCo[3], FlySlip (Gasperin et al., 2007), the Colorado Richly Annotated Full Text (CRAFT) corpus (Cohen et al., 2010b) and DrugNerAr (Segura-Bedmar et al., 2009).

The MEDCo corpus has two subsets, one consisting of abstracts (which we shall refer to as MEDCo-A) and another consisting of full papers (MEDCo-B). The results of our review of all five corpora are presented in Table 1. Included in the last row (HANAPIN) are the attributes of the corpus that we have developed for comparison with existing corpora.

Three of them, MEDSTRACT, MEDCo and DrugNerAr, adapted an annotation scheme similar

[3] http://nlp.i2r.a-star.edu.sg/medco.html

Table 1: Comparison of Biomedical Corpora with Coreference Annotations

Corpus	Scheme Adapted	Document Composition	Domain/ Markables	Coreference Types	Availability	Format	Reliability
MEDSTRACT	MUCCS	100 abstracts	molecular biology/ UMLS types	direct nominal	publicly available	XML	unknown
MEDCo-A	MUCCS	1999 abstracts	human blood cell transcription factors/ GENIA Term Ontology types	direct nominal	publicly available	XML	Krippendorff's alpha: 83% on 15 abstracts
MEDCo-B	MUCCS	43 full papers	human blood cell transcription factors/ GENIA Term Ontology types	direct nominal	currently unavailable	XML	Krippendorff's alpha: 80.7% on 2 full papers
FlySlip	domain-specific	5 full papers	fruit fly genomics/ genetic entities	direct and indirect sortal	publicly available	XML	Kappa score: greater than 83% on each paper
CRAFT	OntoNotes	97 full papers	mouse genomics/ all encountered	direct nominal and verbal and	currently unavailable	SGML	Krippendorff's alpha: 61.9% on 10 full papers
DrugNerAr	MUCCS	49 DrugBank texts	drug-drug interactions/ drugs	direct nominal	publicly available	XML	unknown
HANAPIN	MEDCo	20 full papers	marine natural products chemistry/ chemical compounds, organisms, drug targets, drug effects, diseases	direct nominal, numerical & abbreviation	currently unavailable (to be released publicly)	XML	Krippendorff's alpha: 75% averaged over 20 papers; 84% using the MASI distance metric

to that of the Message Understanding Conference scheme or MUCCS (Hirschman, 1997). Using the Standard Generalized Markup Language (SGML) as annotation format, MUCCS creates a link between co-referring expressions by setting the value of an attribute of the referring element to the ID of the referent.

The same mechanism is used in the annotation of MEDSTRACT, MEDCo and DrugNerAr, but with respective extensions to account for more specific relations (e.g., appositive relation in the case of MEDCo). On the contrary, rather than linking the referring expression to its referent, an annotator explicitly places co-referring expressions in the same coreference chain with OntoNotes, the scheme adapted in annotating the CRAFT corpus. FlySlip can be considered unique in terms of its annotation scheme as it adapted a domain-specific scheme which was necessary since indirect coreferences were annotated. All corpora are available in the form of a mark-up language (SGML or XML).

The five corpora can be grouped into three according to general domain: molecular biology (MEDSTRACT and MEDCo), genomics (FlySlip and CRAFT), and pharmacology (DrugNerAr). MEDSTRACT and MEDCo both have coreference annotations for semantic types from the UMLS and the GENIA ontology, respectively, which can be broadly categorised into compound, organism, protein, gene and cell. Each of the FlySlip and DrugNerAr corpora, on the other hand, have annotations for only one general semantic type: gene-related entities and drugs, respectively. CRAFT is unique in this respect as its developers seek to annotate all co-referring expressions regardless of semantic type; the semantic types that have been encountered so far have not yet been reported, however.

In terms of coreference types for which annotations have been added, CRAFT is the only corpus with annotations for verbal coreference; all the rest have annotations only for pronominal and/or sortal coreference. With respect to coreference types according to relation, FlySlip is the only corpus with annotations for indirect coreference.

MEDCo-B, FlySlip and CRAFT are three existing corpora which are comprised of full-text articles. Among them, only FlySlip is currently publicly available.

The corpus that we have developed, which we call the HANAPIN corpus, is also intended for public release in the near future and covers five general semantic types. In the annotation scheme which was designed and used in HANAPIN, two additional coreference types were considered: abbreviations and numerical coreferences which are commonly used in chemistry research literature. These coreference types and the annotation scheme are further described in the succeeding section.

3 Methodology

3.1 Composition of Corpus Documents

Taking into consideration that the corpus should consist of full-text articles which can be distributed to the public, we gathered full-text articles from the journal *Marine Drugs*[4] which is under the PubMed Central Open Access subset[5]. The said journal covers subject areas such as marine natural products, medicine analysis, marine pharmacology, pharmaceutical biology, marine drugs development and marine biotechnology, among many others. From all of its articles from 2003 to 2009, we randomly selected twenty (20) which seemed to be a reasonable size considering that only five months were allocated for the annotation of the corpus, and that a previous study on biomedical corpora (Cohen et al., 2005) has shown that a corpus can possibly be widely used despite its small size. The experimental sections of the articles were not annotated as they contain very detailed descriptions of the methods carried out by the authors; according to a study (Shah et al., 2003), these usually contain technical data, instruments and measurements – types of information which are currently not of much interest to researchers doing biomedical information extraction, although they may be in the future. The corpus contains a total of 1,027 sentences or 27, 358 words.

3.2 Coreference Types

The coreferences observed in the corpus were categorised into four general nominal types: pronominal, sortal, numerical and abbreviation. Table 2 presents the subtypes of sortal and pronominal coreference, as well as examples for all types. We

[4]http://www.mdpi.com/journal/marinedrugs

[5]http://www.ncbi.nlm.nih.gov/pmc/about/openftlist.html

Table 2: Coreference Types with Examples

General Coreference Type	Subtype	Examples
pronominal	demonstrative	*this, that, these, those*
	personal	*it, they, its, their, theirs*
	indefinite	*another, few, other, some, all, any*
	distributive	*both, such, each, either, neither*
	relative	*which, that, whose*
sortal	definite	*the loihichelins*
	indefinite	*an alkaloid, a mycalamide*
	demonstrative	*this metabolite, these compounds*
	distributive	*both compounds*
	predicate nominative	*"Galactans are **polysaccharides**..."*
	appositive	*"Radiosumin, **an N-methyl dipeptide**..."*
numerical	N.A.	*"The structures of **1** and **2**..."*
		*"Compounds **1-3** inhibit..."*
abbreviation	N.A.	*"...as a membrane type 1 matrix metalloproteinase (MT1-MMP) inhibitor. Compound 1 inhibited **MT1-MMP** with..."*

have decided not to take into account verbal and indirect coreferences; only nominal and direct coreferences have been considered for the first release of the corpus.

3.2.1 Pronominal Coreference

This type of coreference is characterised by a pronoun referring to a noun phrase. The pronoun is used as a substitute to a noun. We have further identified the following subtypes of pronominal coreference: *demonstrative, personal, indefinite, distributive* and *relative*.

3.2.2 Sortal Coreference

Also referred to as lexical noun phrase coreference, sortal coreference is characterised by a noun phrase consisting of a head noun and its modifiers. The subtypes of sortal coreference which have been identified include: *definite, indefinite, demonstrative, distributive, predicate nominative* and *appositive*.

3.2.3 Numerical Coreference

In chemistry research literature, a number is conventionally used to refer to a chemical entity which was introduced using the same number. Oftentimes, a range of numbers is also used to refer to a number of compounds previously mentioned.

3.2.4 Abbreviation

In annotating the HANAPIN corpus, abbreviations were also considered as co-referring expressions. We distinguish them from the other coreference types to make the corpus of benefit to developers of abbreviation identification algorithms as well.

3.3 Annotation Scheme and Procedure

The annotation scheme used in MEDCo (which was based on MUCCS) was adapted and modified for the annotation of the HANAPIN corpus. We have selected the MEDCo scheme as it already differentiates between the pronominal and identity (equivalent to sortal) types, whereas MUCCS has only the identity type. There was a need, however, to extend the MEDCo scheme to further specialise the coreference types. The XML Concordancer (XConc) tool[6] was used in annotating the corpus. Configuring the said tool for our needs is straightforward as it only involved the customisation of a Document Type Definition (DTD) file.

3.3.1 Term Annotations

As a preliminary step, the scheme required that all terms which can be categorised into any of the

[6] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=XConc+Suite

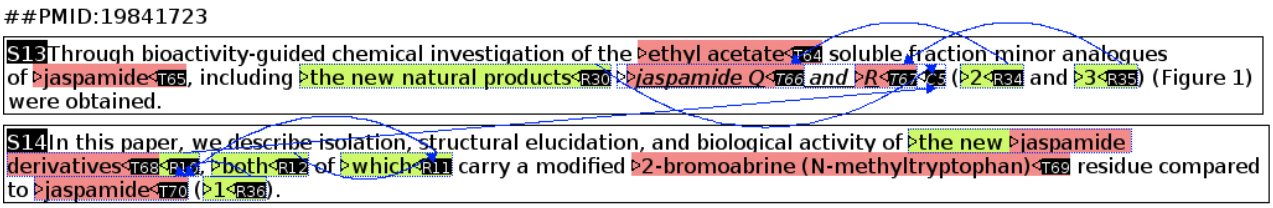

Figure 1: Sample annotations as shown in the XConc annotation tool. The sentences in this example come from one of the documents in the HANAPIN corpus, the *Marine Drugs* article with PubMed ID 19841723. For illustrative purposes, the first sentence in the example was slightly modified to demonstrate the use of the `cons` element.

following semantic types be annotated:

1. chemical compound
2. organism
3. drug effect
4. disease
5. drug target (further categorised into: protein, enzyme, nucleic acid, tissue, cell, cell component, cell line, pathogen)

For each markable, the annotator creates a `term` element which is assigned an ID and one of the semantic types above. The scheme supports the annotation of embedded terms, as well as terms in a discontinuous text region. The former entails placing a `term` element within another. The latter is done by dividing the discontinuous text into fragments and annotating each fragment in the same manner as an ordinary term element. The fragment elements are then grouped together as a constituent element (`cons`). Figure 1 presents a sample annotation of a discontinuous term (constituent `C5`) as viewed in XConc.

3.3.2 Co-referring Expressions

An annotator proceeds to the annotation of co-referring expressions after annotating all terms within a document. If an expression was found to be co-referring with another term, the annotator assigns the ID of the latter as the value of the `idref` attribute of the former. If the referring expression, however, is a noun phrase and not a term that was previously annotated during term annotation, it is marked as a `ref` element and then linked to its referent. Annotators delimit these expressions by including the necessary modifiers of the co-referring element (e.g., *the new jaspamide derivatives* instead of just *jaspamide derivatives*). A coreference type which could be any of pronominal, numerical, abbreviation, and sortal (further categorised into definite, indefinite, demonstrative, distributive, predicate nominative and appositive) is also assigned as the value of the `type` attribute of each link created. We decided not to further divide pronominal coreference into its subtypes as it became apparent during the annotation dry runs that there is only a handful of pronominal coreferences. Figure 1 shows co-referring expressions (connected by arrows) linked by the mechanism just described.

Listed below are some of the main points of the annotation guidelines:

1. A referring expression may be linked to multiple referents.
2. The more specific one between two co-referring expressions is considered as the referent. This means that there might be cases when the referent occurs later than the referring expression. For example, `R30:the new natural products` is the co-referring expression and `C5:jaspamide Q and R` is the referent in Figure 1.
3. In cases where there are multiple choices for the referent of a referring expression, the closest one may be chosen as long as it is (or will be) linked to the other choice expressions.
4. There are cases when more than one type of coreference applies. For example, in Figure 1, *the new natural products* is both an appositive and a definite noun phrase. In such cases, the appositive and predicate nominative types take precedence over the other sortal types.

```
<sentence id="S13">
    Through bioactivity-guided chemical investigation of the
    <term id="T64" sem="chem">ethyl acetate</term>
    soluble fraction minor analogues of
    <term id="T65" sem="chem">jaspamide</term>, including
    <ref id="R30" idref1="C5" type="appos">the new natural products</ref>
    <cons id="C5">
        <term id="T66" sem="chem">jaspamide Q</term> and
        <term id="T67" sem="chem">R</term>
    </cons> (
    <ref id="R34" idref1="T66" type="num">2</ref> and
    <ref id="R35" idref1="T67" type="num">3</ref>) (Figure 1) were obtained.
</sentence>
<sentence id="S14">In this paper, we describe isolation, structural elucidation,
    and biological activity of
    <ref id="R10" idref1="C5" type="definite">the new
        <term id="T68" sem="chem">jaspamide derivatives</term>
    </ref>,
    <ref id="R12" idref1="R11" type="pron">both</ref> of
    <ref id="R11" idref1="R10" type="pron">which</ref> carry a modified
    <term id="T69" sem="chem">2-bromoabrine (N-methyltryptophan)</term>
    residue compared to
    <term id="T70" sem="chem">jaspamide</term> (
    <ref id="R36" idref1="T70" type="num">1</ref>).
</sentence>
```

Figure 2: XML code generated by XConc for the sample annotations in Figure 1.

One could process the XML code (provided in Figure 2 for the reader's reference) to obtain the following coreference chains:

1. {R30:the new natural products, C5:jaspamide Q and R, R10:the new jaspamide derivatives, R11:which, R12:both}
2. {T66:jaspamide Q, R34:2}
3. {T67:jaspamide R, R35:3}
4. {T70:jaspamide, R36:1}

The complete annotation guidelines will be publicly released together with the annotated corpus.

4 Results

The three annotators were asked to complete the coreference annotations within five months. A biweekly meeting was held to address questions and issues which could not be addressed or resolved by means of the online project forum.

4.1 Statistics

As the HANAPIN corpus is the first of its kind from the biochemistry domain and aims to cover several semantic as well as coreference types, it is of interest to determine which of the types are most prevalent. To do this we computed statistics over the annotations (Figure 3). For each type, we obtained the average over the annotations from the three coders.

There is a total of 395 coreference chains (not including singleton chains or those with only one mention) in the entire corpus. The coreference chains are of the following semantic types: chemical compounds (70.89%), drug targets (12.66% that accounts for proteins, cell lines, pathogens, enzymes, cells, cell parts, nucleic acids and tissues), organisms (9.87%), drug effects (3.29%), and diseases (3.29%). Among the drug targets, the most prevalent are proteins, cell lines and pathogens.

A total of 760 coreference links have been found in the corpus. The most common among the types is the numerical one (46%), followed by the sortal type (33% that accounts for the definite, indefinite, demonstrative, appositive, predicate nominative and distributive types). Less common are the pronominal type (11%) and abbreviation (10%). Among the sortal coreferences, the most common are the definite and indefinite types, followed by the demonstrative type.

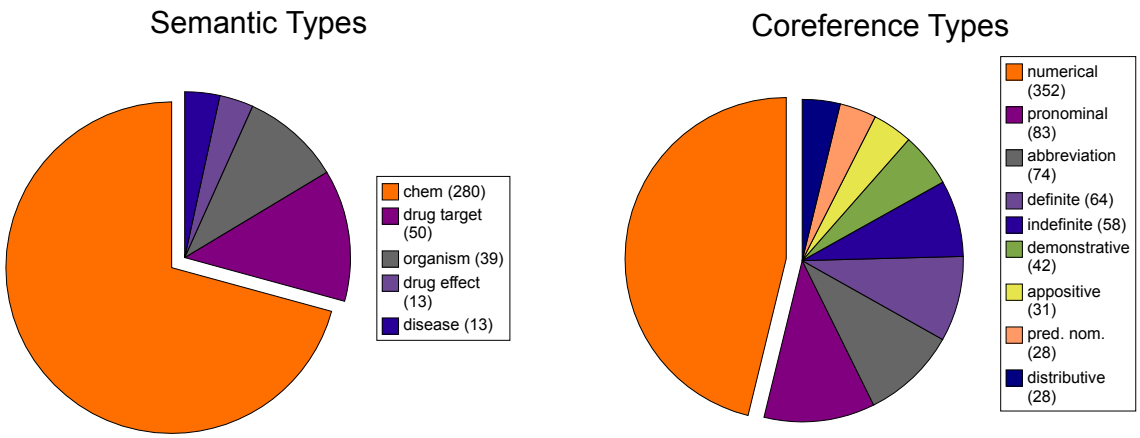

Figure 3: Distribution of semantic and coreference types in the HANAPIN corpus.

4.2 Corpus Reliability

Following Passoneau's proposed method for computing reliability for coreference annotation (Passoneau, 2004), we computed for the reliability of the corpus in terms of Krippendorff's alpha, a coefficient of agreement that allows for partial disagreement with the use of a distance metric based on the similarity between coreference chains. Passoneau's first proposed distance metric (d_P) assigns 0 for identity, 0.33 for subsumption, 0.67 for intersection and 1 for disjunction. There are, however, alternative distance metrics that consider the sizes of the coreference chains, such as Jaccard's coefficient of community (d_J) and Dice's coincidence index (d_D) which can be computed as follows (Artstein and Peosio, 2004):

$$d_J = 1 - \frac{|A \cap B|}{|A \cup B|}$$

$$d_D = 1 - \frac{2|A \cap B|}{|A| + |B|}$$

A new distance metric called Measuring Agreement on Set-valued Items (MASI) was then later proposed by Passoneau. It is obtained by getting the product of the original distance metric d_P and Jaccard's coefficient d_J.

Initially using Passoneau's first proposed distance metric d_P in computing for Krippendorff's alpha, we obtained an average of 75% over all documents in the HANAPIN corpus. Computing for alpha using the MASI distance metric gives 84%. Though there is no value of alpha that has been established to be an absolute indication of high agreement, previous works cited by Krippendorff have shown that values of alpha less than 67% indicate unreliability (Krippendorff, 1980). We can therefore regard the obtained values of alpha as satisfactory.

5 Conclusion and Future Work

A coreference-annotated corpus from the domain of biochemistry, consisting of full-text articles, has been developed. It was annotated following guidelines which covered coreference and semantic types that have not been covered in other biomedical corpora before. This was done to further the aim of providing researchers with more insight into the phenomenon of coreference in a cross-disciplinary domain. Results show that in this biochemistry domain, the most common types of coreference being used by authors are the numerical and sortal types. Verbal and indirect coreferences, however, have not been considered at this stage; the annotation of these types can be explored as part of future work on the corpus.

To measure reliability of the corpus, we determined inter-annotator agreement on all documents by computing for the value of Krippendorff's alpha. Using Passoneau's first proposed distance metric and the MASI distance metric, we obtained satisfactory values of 75% and 84%, respectively. The corpus and annotation guidelines will be released to the public to encourage and enable more researchers to develop improved biomedical coreference resolu-

tion methodologies.

Acknowledgements

The UK National Centre for Text Mining is funded by the UK Joint Information Systems Committee (JISC). The authors would also like to acknowledge the Office of the Chancellor, in collaboration with the Office of the Vice-Chancellor for Research and Development, of the University of the Philippines Diliman for funding support through the Outright Research Grant.

The authors also thank Paul Thompson for his feedback on the annotation guidelines, and the anonymous reviewers for their helpful comments.

References

Ron Artstein and Massimo Poesio. 2004. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555-596.

José Castaño, Jason Zhang and James Pustejovsky. 2002. Anaphora resolution in biomedical literature. *Proceedings of the International Symposium on Reference Resolution for NLP*.

K. Bretonnel Cohen, Philip V. Ogren, Lynne Fox and Lawrence E. Hunter. 2005. Empirical data on corpus design and usage in biomedical natural language processing. *AMIA Annual Symposium Proceedings*, pages 156-160.

K. Bretonnel Cohen, Helen L. Johnson, Karin Verspoor, Christophe Roeder, Lawrence E. Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11(1):492.

K. Bretonnel Cohen, Arrick Lanfranchi, William Corvey, William A. Baumgartner Jr., Christophe Roeder, Philip V. Ogren, Martha Palmer and Lawrence E. Hunter. 2010. Annotation of all coreference in biomedical text: Guideline selection and adaptation. *Proceedings of the Second Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2010), LREC 2010*.

Caroline Gasperin, Nikiforos Karamanis and Ruth Seal. 2007. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2007)*.

Ralph Grishman and Beth Sundheim. 1995. Design of the MUC-6 Evaluation. *MUC '95: Proceedings of the 6th Message Understanding Conference*, pages 1-11.

Lynette Hirschman. 1997. MUC-7 Coreference Task Definition. *Message Understanding Conference 7 Proceedings*.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, 2nd edition.

Klaus H. Krippendorff. 1980. Content Analysis: An Introduction to Its Methodology. Beverly Hills, CA: Sage Publications.

Tara McIntosh and James R.Curran. 2009. Challenges for automatically extracting molecular interactions from full-text articles. *BMC Bioinformatics*, 10(1):311.

Ruslan Mitkov, Richard Evans, Constantin Orasan, Catalina Barbu, Lisa Jones and Violeta Sotirova. 2005. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC 2000)*, pages 49-58.

Rebecca J. Passoneau. 2004. Computing reliability for coreference annotation. *Proceedings of the International Conference on Language Resouces (LREC)*.

M. Schumie, M. Weeber, B. Schijvenaars, E. van Mulligen, C. van der Eijk, R. Jelier, B. Mons and J. Kors. 2004. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16):2597-2604.

Isabel Segura-Bedmar, Mario Crespo, César de Pablo-Sánchez and Paloma Martínez. 2009. Resolving anaphoras for the extraction of drug- drug interactions in pharmacological documents. *BMC Bioinformatics*, 11(Suppl 2):S1.

Parantu K. Shah, Carolina Perez-Iratxeta, Peer Bork and Miguel A. Andrade. 2003. Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics*, 4(1): 20.