BioAR: Anaphora Resolution for Relating Protein Names to Proteome Database Entries

Jung-jae Kim[†] **and Jong C. Park**[‡]
Computer Science Division & AITrc
Korea Advanced Institute of Science and Technology
373-1, Guseong-dong, Yuseong-gu, Daejeon 305-701, South Korea
[†]jjkim@nlp.kaist.ac.kr
[‡]park@cs.kaist.ac.kr

Abstract

The need for associating, or grounding, protein names in the literature with the entries of proteome databases such as Swiss-Prot is well-recognized. The protein names in the biomedical literature show a high degree of morphological and syntactic variations, and various anaphoric expressions including null anaphors. We present a biomedical anaphora resolution system, BioAR, in order to address the variations of protein names and to further associate them with Swiss-Prot entries as the actual entities in the world. The system shows the performance of 59.5%∼75.0% precision and 40.7%∼56.3% recall, depending on the specific types of anaphoric expressions. We apply BioAR to the protein names in the biological interactions as extracted by our biomedical information extraction system, or BioIE, in order to construct protein pathways automatically.

1 Introduction

The need for identifying the antecedents of anaphoric expressions in the literature is well-recognized. Most previous approaches assume that anaphoric expressions and their antecedents would appear *in the same documents*. However, further work is called for when such antecedents need to be associated with actual entities in the world, where the task of establishing the denotation of a named entity with respect to the world or a model is known as named entity grounding (Leidner et al., 2003). In the biomedical domain where the phrases in the literature tend to refer to actual biological entities such as proteins, the phrases should be associated with the actual entries of external resources (Hachey et al., 2004). In this paper, we present a biomedical anaphora resolution system, BioAR, in order to identify the actual referents of those phrases in the biomedical literature and to annotate the phrases, especially those that refer to proteins, with the entries of proteome databases such as Swiss-Prot by suitable anaphora resolution.

Anaphora resolution indicates the process of determining the antecedent of an anaphoric expression. Traditional approaches to anaphora resolution in general domain utilize various constraints or preferences from the morphological, syntactic, and semantic points of view. The most prominent proposal for anaphora resolution is a centering theory (Grosz et al. (1995)), which identifies the antecedents of pronouns with respect to discourse structures, based on the observation that those entities that have already been mentioned and are more central than others tend to be referred back by pronouns subsequently. Byron (2002) proposed to identify the antecedents of pronominal references in spoken dialogues by utilizing discourse structures with discourse entities and semantic filtering. Castano et al. (2002) adopted a knowledge-poor method, which focuses on resolving pronouns robustly, for example with part-of-speech information, positions of the candidate antecedents, agree-

(1)	The yeast and mammalian branchpoint sequence binding proteins (BBP and mBBP/SF1) contain both KH domain and Zn knuckle RNA-binding motifs. ··· Therefore, we propose that all three of these accessory RNA-binding modules bind the phosphate backbone, whereas *the KH domain* interacts specifically with the bases of the BPS. (PMID:9701290)

Table 1: A protein domain-referring phrase example

(2)	MOB1 exhibits genetic *interaction* with three other yeast genes required for the completion of mitosis, LTE1, CDC5, and CDC15 (the latter two encode essential protein kinases). (PMID:9436989)
(3)	Screening for the emerin binding protein and immunopr ecipitation analysis showed that lamin A binds to emerin specifically. We also used the yeast two-hybrid system to clarify that *this interaction* requires the top half of the tail domain (amino acid 384-566) of lamin A. (PMID:11173535)

Table 2: Missing argument examples of biological interactions

ments and lexical features, in addressing problems in the biomedical domain (cf. Mitkov et al. (1998)).

In the biomedical literature, an anaphoric expression works as the device of making an abbreviated and indirect reference to some biological object or objects. This notion can be applied to all the phrases in the literature which refer to proteins, in that the phrases can be associated (or grounded) with the protein entries in proteome databases, which biologists generally regard as the identities of proteins. The protein-referring phrases in the literature include not only gene/protein names but also anaphoric expressions and missing arguments of biological interactions (or null anaphors) which refer to proteins.[1]

As for anaphoric expressions, previous approaches to anaphora resolution often stop at antecedent noun phrases in the same documents, but we propose to further identify the proteins that are composed of protein domains referred to by anaphoric expressions. For example, the anaphoric expression *the KH domain* in the last sentence in Table 1 refers to the domain shared by the proteins "the yeast and mammalian branchpoint sequence binding proteins (BBP and mBBP/SF1)."[2]

While previous approaches have dealt only with the resolution of pronouns (e.g. *it*, *they*) and sortal anaphoric noun phrases (e.g. *the protein*, *both enzymes*), we can also restore the missing arguments of the biological interactions, mostly represented with nominal interaction keywords such as *interaction* with or without determiners, by utilizing the context (cf. Hong and Park (2004)). For example, the omitted argument of *interaction* in the first example in Table 2 is the sentential subject, or "MOB1." In the second example in Table 2, the two omitted participants of the interaction represented by the anaphoric expression *this interaction* are "lamin A" and "emerin," which are also the syntactic arguments of the verb *binds*.

In this paper, we present a biomedical anaphora resolution system, BioAR, to ground the protein-referring phrases in the biological interactions extracted by our biomedical information extraction system, BioIE (Kim and Park, 2004; Park et al., 2001), with Swiss-Prot entries. BioIE is a system that extracts general biological interactions of *arbitrary types* from the biomedical literature. This system shows the performance of 88∼92% precision and 55∼57% recall, or the F-scores of 68∼70. While the output of BioIE includes complex linguistic phenomena, such as anaphoric expressions, conjunctions, prepositional phrases, and relative clauses, many of the noun phrases in the results of BioIE refer to proteins since the relevant interaction keywords, such as *interact* and *bind*, mostly represent protein-protein interactions

[1] As for anaphora resolution, there are three related kinds of objects in biomedical domain, that is, pronouns, antecedents, and real entities in the world, where pronouns and antecedents are the phrases in the literature. Among antecedents, there can be "anaphoric" ones, referring to other antecedents. Both pronouns and antecedents eventually refer to real entities in the world, so the protein-referring phrases in the literature include both pronouns and antecedents in the literature.

[2] Hereafter, the italicized string is an anaphoric expression and the underlined string is its antecedent.

Anaphoric expression	Count
Pronouns	53
Anaphoric DNPs	26
Missing arguments	8

Table 3: Statistics of anaphoric expressions

and the interactions among them.[3] BioAR grounds those protein-referring phrases with Swiss-Prot entries which work as the protein nodes in the protein pathways that can be automatically built by incorporating the biological interactions extracted by BioIE.

2 Methods

BioAR identifies the antecedents of anaphoric expressions that appear in the results of BioIE and annotates the protein-referring phrases with Swiss-Prot entries. The system first locates pronouns, noun phrases with determiners (DNPs), and biological interactions as the candidates of anaphoric expressions. Table 3 shows the statistics of these anaphoric expressions.[4] The rest of the system is implemented in the following four steps: 1) pronoun resolution, 2) resolution of anaphoric DNPs, 3) restoration of missing arguments in the biological interactions, and 4) grounding the protein-referring phrases with Swiss-Prot entries.

2.1 Pronoun resolution

We adopt the centering theory of Grosz et al. (1995) for the anaphora resolution of pronouns. In particular, we follow the observation that the entities which have already been mentioned and are more central than others tend to be referred back by pronouns subsequently. For example, the candidate antecedent in the sentential subject is preferred to that in the sentential object (cf. Table 4).

As for possessive pronouns such as *its* and *their*, we have found that the antecedents of these possessive pronouns are mostly located in the same or preceding sentences and that possessive pronouns can be classified into the following two types according to the sentential locations of their antecedents,

[3]There are 232 noun phrases which can be associated with Swiss-Prot entries, among 1,645 noun phrases in 516 biological interactions extracted by BioIE from a subset of *yeast* corpus.

[4]We have counted the anaphoric expressions among 1,645 noun phrases in the subset of *yeast* corpus.

(4)	Finally, SpNAC can bind to X-junctions that are already bound by a tetramer of the Escherichia coli RuvA protein, indicating that *it* interacts with only one face of the junction. (PMID:11243781)

Table 4: A subjective pronoun resolution example

where 1) the antecedent of a possessive pronoun is the protein name which is nearest to the left of the possessive pronoun in the same sentence and 2) the antecedent of another possessive pronoun is the leftmost protein name in the subject phrase of the same or preceding sentence (cf. Table 5). We have also found that the local context of a possessive pronoun of the second type mostly shows syntactic parallelism with that of its antecedent, as in the two *they* of the second example in Table 5, while that of the first type does not show parallelism where the antecedents of such possessive pronouns are mostly the protein names nearest to the left of the possessive pronouns.[5] Since the antecedents of possessive pronouns of the second type can be detected with the patterns that encode the parallelism between the local context of a possessive pronoun and that of its antecedent in the same sentence (cf. Table 6),[6] we have set the protein names, those nearest to the left of the possessive pronouns in the same sentences, as the default antecedents of possessive pronouns and utilized the patterns, such as those in Table 6, in recognizing the possessive pronouns of the second type and in locating their antecedents.

2.2 Noun phrase resolution

In the process of resolving anaphoric noun phrases, BioAR first locates the noun phrases with determiners (DNPs), especially those with definites (i.e. *the*) and demonstratives (i.e. *this*, *these*, and *those*), as

[5]Among the 1,000 biological interactions, there are 31 possessive pronouns of the first type and 17 possessive pronouns of the second type.

[6]POSS indicates a possessive pronoun; ANT indicates its antecedent; NP which follows POSS indicates the rest of the noun phrase which starts with POSS; and BeV indicates a be-verb. VB, VBN, and PP are `POS` tags, indicating main verbs, past particles, and prepositions, respectively. 'A|B' indicates that either A or B should occur. '$\cdots$' can be matched to any sequence of words.

(5)	Using the Yeast Two-Hybrid system and further in vitro and in vivo studies, we identified the regulatory beta-subunit of casein kinase II (CKII), which specifically binds to the cytoplasmic domain of CD163 and *its* isoforms. (PMID:11298324)
(6)	F-box proteins are the substrate-recognition components of SCF (Skp1-Cullin-F-box protein) ubiquitin-protein ligases. *They* bind the SCF constant catalytic core by means of the F-box motif interacting with Skp1, and *they* bind substrates through *their* variable protein-protein interaction domains. (PMID:11099048)

Table 5: Possessive pronoun resolution examples

1. via|through|due to POSS NP
2. ANT BeV VBN · · · and VBN PP POSS NP
3. ANT BeV VBN and POSS NP VBN PP
4. ANT BeV VBN · · · and POSS NP BeV VBN
5. VB that ANT VB · · · ,|and that POSS NP
6. ANT VB · · · , and POSS NP VB
7. ANT's NP VB · · · and POSS NP VB

Table 6: Example patterns for parallelism

the candidates of anaphoric noun phrases.[7] Among the noun phrases with definites, the noun phrases that do not have antecedents in the context, i.e. non-anaphoric DNPs, mostly belong to the classes in Table 7.[8] [9] The system filters out those non-anaphoric DNPs belonging to those classes in Table 7, by utilizing a list of cellular component names, a list of species names, and the patterns in Table 7 which represent the internal structures of some non-anaphoric DNPs. We have also developed modules to identify appositions and acronyms in order to filter out remaining non-anaphoric DNPs.

BioAR scores each candidate antecedent of an anaphoric DNP with various salience measures and identifies the candidate antecedent with the highest score as the antecedent of the anaphoric DNP (cf. Castano et al. (2002)). For example, the system assigns penalties to the candidate antecedents whose numbers do not agree with those of anaphoric DNPs. Among the candidate antecedents of anaphoric DNPs, the candidate antecedents in the sentential subjects are preferred to those in the sentential objects or other noun phrases, following the centering theory (Grosz et al., 1995). We have also adopted salience measures to score each candidate antecedent according to the morphological, syntactic, and semantic characteristics of candidate antecedents (cf. Castano et al. (2002)). For example, when a DNP refers to a protein, its candidate antecedents which refer to protein domains get negative scores, and when a DNP refers to a protein domain, its candidate antecedents which refer to protein domains get positive scores. Furthermore, when a DNP refers to an enzyme, its candidate antecedents which end with '-ase' get positive scores.

In the process of resolving the anaphoric DNPs referring to protein domains, the system identifies the proteins which contain the domains referred to by the anaphoric expressions. We have constructed several syntactic patterns which describe the rela-

1.	(39)	DNP modified by a prepositional phrase or a relative clause (Ex. *the C-terminal of AF9*)
2.	(24)	DNP of the pattern 'DEFINITE GENE protein' (Ex. *the E6 protein*)
3.	(16)	DNP with appositive structure (Ex. *the yeast transcriptional activator Gcn4*)
4.	(10)	DNP ending with acronyms (Ex. *the retinoid X receptor (RXR)*)
5.	(6)	DNP of the pattern 'DEFINITE PROTEIN DOMAIN' (Ex. *the DNA-PK catalytic subunit*)
6.	(4)	DNP indicating a cellular component (Ex. *the nucleus*)
7.	(2)	DNP indicating a species name (Ex. *the yeast Saccharomyces cerevisiae*)

Table 7: Non-anaphoric DNP examples

[7] We also deal with other anaphoric noun phrases with 'both' or 'either', as in 'both proteins' and 'either protein'.

[8] GENE, PROTEIN, and DOMAIN indicate a gene name, a protein name, and a generic term indicating protein domain such as *domain* and *subunit*, respectively. DEFINITE indicates the definite article *the*.

[9] The digit in parentheses indicates the number of non-anaphoric DNPs in each class, among 117 DNPs in 390 biological interactions.

1. DOMAIN of|in PROTEIN
2. PROTEIN BeV NN composed of DOMAIN
3. PROTEIN BeV NN comprising DOMAIN
4. PROTEIN contain DOMAIN
5. the PROTEIN DOMAIN

Table 8: Example patterns of proteins and their domains

tionships between proteins and their domains as exemplified in Table 8.

The system locates the coordinate noun phrases with conjunction items such as 'and', 'or', and 'as well as' as the candidate antecedents of plural anaphoric expressions. The system also locates the proteins in the same protein family in the same document, as in *MEK1 and MEK2*, as the candidate antecedent of a plural anaphoric expression such as *these MEKs* (PMID:11134045).

2.3 Biological interaction resolution

BioAR also restores some of the missing arguments of interaction keywords by utilizing the context. When one or more syntactic arguments of biological interactions in the results of BioIE are elided, it is essential to identify the antecedents of the omitted arguments of the interactions, or null anaphora, as well. We have focused on resolving the missing arguments of nominal interaction keywords, such as *interaction*, *association*, *binding*, and *co-immunoprecipitate*,[10] based on the observation that those keywords mostly represent protein-protein interactions, and thus their omitted arguments refer to proteins or protein domains in the previous context. In case only one argument of an interaction keyword is elided as in the first example in Table 2, the proteins in the sentential subjects are preferred as antecedents to those in other noun phrases of the sentences which contain the interaction keyword. In case both arguments of an interaction keyword are elided as in the second example in Table 2, both the sentences, whose main verbs are in the verbal form

[10]The interaction keywords of interest, *interaction*, *association*, *binding*, and *co-immunoprecipitate*, indicate physical binding between two proteins, and thus they can be replaced with one another. In addition to them, the interaction keywords *phosphorylation* and *translocation* also often indicate protein-protein interactions.

1. interaction of A with B
2. association of A with B
3. co-immunoprecipitation of A with B
4. binding of A to B
5. interaction between|among A and B
6. association between|among A and B
7. co-immunoprecipitation between|among A and B
8. binding between|among A and B

Table 9: Example patterns of nominal interaction keywords

(7)	Interactions among the three MADS domain proteins were confirmed by in vitro experiments using GST-fused OsMADS1 expressed in Escherichia coli and in vitro translated proteins of OsMADS14 and -15. ⋯ While the K domain was essential for protein-protein interaction, a region preceded by the K domain augmented *this interaction*. (PMID:11197326)

Table 10: An example antecedent of a nominal interaction keyword

of the interaction keyword, and the noun phrases of the patterns in Table 9, whose headwords are the same as the interaction keyword, can be the candidate antecedents of the interaction keyword with its two missing arguments. Table 10 shows an example antecedent with a nominal interaction keyword.

2.4 Protein name grounding

We have constructed around 0.7 million gene and protein names from the gene name (GN) and description (DE) fields of Swiss-Prot in order to recognize protein names in the literature. We have also developed several patterns to deal with the variations of protein names (cf. Table 11). Table 12 shows several examples of grounding protein names with Swiss-Prot entries.[11]

Taking into account the fact that many Swiss-Prot entries actually indicate certain domains of bigger proteins, for example *Casein kinase II beta chain* (KC2B_YEAST) and *Ribonuclease P protein*

[11]The terms of the form A_B, where B indicates the species information, are Swiss-Prot entries.

Swiss-Prot term	Variation
D(2)	D2
S-receptor kinase	S receptor kinase
RNase P protein	RNase P
Thioredoxin h-type 1	Thioredoxin h (THL1)

Table 11: Term variation examples

Protein name	Swiss-Prot entries
Filamin A	FLNA_HUMAN, FLNA_MOUSE
Pop1p	POP1_HUMAN, POP1_SCHPO, POP1_YEAST
D3 dopamine receptor	D3DR_CERAE, D3DR_HUMAN, D3DR_MOUSE, D3DR_RAT

Table 12: Protein name grounding examples

component (RPM2_YEAST), BioAR grounds the phrases in the results of BioIE, which refer to protein domains, with the descriptions of Swiss-Prot entries, by converting those phrases into the structures as utilized by Swiss-Prot. For example, the phrase "the regulatory beta-subunit of casein kinase II (CKII)" can be grounded with KC2B_YEAST, and the phrase "the individual protein subunits of eukaryotic RNase P" with RPM2_YEAST. Furthermore, the information about the domains of a protein is sometimes described in the SUBUNIT field of Swiss-Prot. For example, the protein domain name "the RNA subunit of RNase P" can be grounded with *RPM1* in the SUBUNIT field of RPM2_YEAST, i.e. "Consists of a RNA moiety (RPM1) and the protein component (RPM2). Both are necessary for full enzymatic activity." We leave the problem of looking up the SUBUNIT field of Swiss-Prot as future work.

Since a protein name can be grounded with multiple Swiss-Prot entries as shown in Table 12, BioAR tries to choose only one Swiss-Prot entry, the most appropriate one for the protein name among the candidate entries, by identifying the species of the protein from the context (cf. Hachey et al. (2004)). For example, while the protein name *Rpg1p/Tif32p* can be grounded with two Swiss-Prot entries, or {IF3A_SCHPO, IF3A_YEAST}, the noun phrase "Saccharomyces cerevisiae Rpg1p/Tif32p" should be grounded only with IF3A_YEAST. Similarly, the system grounds the protein name *Sla2p*

(8) The yeast two-hybrid system was used to screen for proteins that interact in vivo with Saccharomyces cerevisiae Rpg1p/Tif32p, the large subunit of the translation initiation factor 3 core complex (eIF3). Eight positive clones encoding portions of the *SLA2*/END4/MOP2 gene were isolated. Subsequent deletion analysis of *Sla2p* showed that amino acids 318-373 were essential for the two-hybrid protein-protein interaction. (PMID:11302750)

Table 13: An annotation example for the necessity of species information

only with SLA2_YEAST among candidate Swiss-Prot entries, or {SLA2_HUMAN, SLA2_MOUSE, SLA2_YEAST}, when the protein name occurs together with the species name *Saccharomyces cerevisiae* in the same abstract as in Table 13.

In summary, BioAR first locates anaphoric noun phrases, such as pronouns and anaphoric DNPs, and interaction keywords that appear in the results of BioIE, while it filters out non-anaphoric DNPs and the interaction keywords with two explicit syntactic arguments. The system identifies the antecedents of pronouns by utilizing patterns for parallelism and by following the observation in the centering theory. The system identifies the antecedents of anaphoric DNPs by utilizing various salience measures. In particular, the system identifies the proteins which contain the protein domains referred to by anaphoric expressions. The system restores the missing arguments of biological interactions from the context. Finally, the system grounds the protein-referring phrases in the results of BioIE with the most appropriate Swiss-Prot entry or entries.

3 Experimental results

We have developed BioAR with a training corpus consisting of 7,570 biological interactions that are extracted by BioIE from 1,505 MEDLINE abstracts on *yeast* (cf. Kim and Park (2004)). BioAR takes 24 seconds to process 1,645 biological interactions in the training corpus. We have constructed a test corpus which is extracted from MEDLINE with a different MeSH term, or *topoisomerase inhibitors*.

SOURCE	
PMID	10022855
Sentence	Gadd45 could potentially mediate this effect by destabilizing histone-DNA interactions since *it* was found to **interact** directly with *the four core histones*.
INTERACTION	
Keyword	**interact**
Argument1	*it*
Argument2	*the four core histones*
PRONOUN RESOLUTION	
Anaphor	*it*
Antecedent	Gadd45
PROTEIN NAME GROUNDING	
Phrase	Gadd45
S-P entry	GA45_HUMAN

Table 14: An example result of BioAR

	Precision	Recall
Pronoun resolution	75.0% (9/12)	56.3% (9/16)
Noun phrase resolution	75.0% (12/16)	52.2% (12/23)
Protein name grounding	59.5% (22/37)	40.7% (22/54)

Table 15: Experimental results of test corpus

The test corpus includes 120 unseen biological interactions extracted by BioIE. Table 15 shows the experimental results of the modules of BioAR on the test corpus.[12] Table 14 shows an example result of BioAR.

4 Discussion

We have analyzed the errors from each module of BioAR. All the incorrect antecedents of pronouns

[12]While the missing arguments of biological interactions often occur in the training corpus, there was only one missing argument in the test corpus, which is correctly restored by BioAR. This result is included into those of noun phrase resolution. Moreover, the rules and patterns utilized by BioAR show a low coverage in the test corpus. It would be helpful to utilize a machine-learning method to construct such rules and patterns from the training corpus, though there are few available anaphora-tagged corpora.

(10)	These triterpenoids were not only mammalian DNA polymerase inhibitors but also inhibitors of DNA topoisomerases I and II even though the enzymic characteristics of DNA polymerases and DNA topoisomerases, including their modes of action, amino acid sequences and three-dimensional structures, differed markedly. ... Because the three-dimensional structures of fomitellic acids were shown by computer simulation to be very similar to that of ursolic acid, the DNA-binding sites of both enzymes, which compete for *the inhibitors*, might be very similar. (PMID:10970789)

Table 16: Incorrect resolution example of pronoun resolution module

in the test corpus produced by the pronoun resolution module are due to incorrect named entity recognition, as in the incorrectly identified named entity "DNA double-strand" from the phrase "DNA double-strand break (DSB)" and "-II' in "topo-I or -II." This problem can be dealt with by a domain-specific POS tagger and a named entity recognizer. Further semantic analysis with the help of the context is needed to deal with the errors of noun phrase resolution module. For example, "these triterpenoids" in Table 16 are *inhibitors*, and thus it can be a candidate antecedent of the anaphoric DNP *the inhibitors*.

In the process of protein name grounding, BioAR grounds 8 abbreviations among 15 incorrectly grounded protein-referring phrases with irrelevant Swiss-Prot entries. Furthermore, among 32 protein-referring phrases not grounded by BioAR, 14 phrases are the same as the string *topoisomerase* where the string always indicates "DNA topoisomerase" in the corpus of *topoisomerase inhibitors*. To address this problem, we need domain-specific knowledge, which we leave as future work.

Castano et al. (2002) presented a knowledge-poor method to utilize salience measures, including parts-of-speech, positions of the candidate antecedents, agreements and lexical features. While the method reportedly shows a relatively high performance of

77% precision and 71% recall, we note that the method is unable to deal with domain-specific anaphora resolution, for example the task of identifying the proteins which contain the protein domains referred to by anaphoric expressions.

Leidner et al. (2003) presented the method of grounding spatial named entities by utilizing two minimality heuristics, that is, that of assuming one referent per discourse and that of selecting the smallest bounding region in geographical maps. Hachey et al. (2004) presented a method for grounding gene names with respect to gene database identifiers by dealing with various kinds of term variations and by removing incorrect candidate identifiers with statistical methods and heuristics. These methods are similar to BioAR in that they also aim to ground the phrases in texts with respect to the entities in the real world. However, BioAR further contributes to biomedical named entity grounding by dealing with the relationships between proteins and their domains and by identifying the species information of protein names from the context.

5 Conclusion

BioAR identifies the antecedents of anaphoric noun phrases that appear in the results of BioIE. The system further identifies the proteins which contain the domains referred to by anaphoric expressions by utilizing several patterns which describe their relations. The system also identifies the missing arguments of biological interactions by utilizing biological interaction patterns. Finally, the system grounds the protein-referring phrases with the most relevant Swiss-Prot entries by consulting the species information of the proteins.

We believe that anaphora resolution with database entries may not be addressed in other domains as straightforwardly as in this paper, since there are quite few comprehensive resources with actual entities. The task of grounding the protein-referring phrases in the results of BioIE with Swiss-Prot entries is crucial to building up incorporated protein pathways consisting of the biological interactions extracted by BioIE. We are currently working on integrating BioIE, BioAR, and other systems for ontology manipulation and information visualization for synergistic knowledge discovery.

Acknowledgement

We are grateful to the anonymous reviewers and to Bonnie Webber for helpful comments. This work has been supported by the Korea Science and Engineering Foundation through AITrc.

References

Byron, D.K. 2002. Resolving pronominal reference to abstract entities. *Proc. ACL*, 80–87.

Castano, J., Zhang, J., and Pustejovsky, J. 2002. Anaphora resolution in biomedical literature. *Int'l Symp. Reference Resolution in NLP*, Alicante, Spain.

Grosz, B.J., Joshi, A.K., and Weinstein, S. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 203–225.

Hachey, B., Nguyen, H., Nissim, M., Alex, B., and Grover, C. 2004. Grounding gene mentions with respect to gene database identifiers. *Proc. the BioCreative Workshop*, Granada, Spain.

Hong, K.W. and Park, J.C. 2004. Anaphora Resolution in Text Animation. *Proc. the IASTED International Conference on Artificial Intelligence and Applications (AIA)*, pp. 347–352, Innsbruck, Austria.

Kim, J.-J. and Park, J.C. 2004. BioIE: retargetable information extraction and ontological annotation of biological interactions from the literature. *J. Bioinformatics and Computational Biology*. (to appear)

Leidner, J.L., Sinclair, G., and Webber, B. 2003. Grounding spatial named entities for information extraction and question answering. *Proc. the HLT/NAACL'03 Workshop on the Analysis of Geographic References*, Edmonton, Alberta, Canada, May.

Mitkov, Rulsan. 1998. Robust pronoun resolution with limited knowledge. *Proc. COLING/ACL*, 869–875.

Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. 2001. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161.

Park, J.C., Kim, H.S., and Kim, J.J. 2001. Bidirectional incremental parsing for automatic pathway identification with Combinatory Categorial Grammar. *Proc. PSB*, 396–407.