Relation Annotation for Understanding Research Papers

Yuka Tateisi† Yo Shidahara‡ Yusuke Miyao† Akiko Aizawa†
†National Institute of Informatics, Tokyo, Japan
{yucca,yusuke,aizawa}@nii.ac.jp
‡Freelance Annotator
yo.shidahara@gmail.com

Abstract

We describe a new annotation scheme for formalizing relation structures in research papers. The scheme has been developed through the investigation of computer science papers. Using the scheme, we are building a Japanese corpus to help develop information extraction systems for digital libraries. We report on the outline of the annotation scheme and on annotation experiments conducted on research abstracts from the IPSJ Journal.

1 Introduction

Present day researchers need services for searching research papers. Search engines and publishing companies provide specialized search services, such as Google Scholar, Microsoft Academic Search, and Science Direct. Academic societies provide archives of journal articles and/or conference proceedings such as the ACL Anthology. These services focus on simple keyword-based searches as well as extralinguistic relations among research papers, authors, and research topics. However, because contemporary research is becoming increasingly complicated and interrelated, intelligent content-based search systems are desired (Banchs, 2012). A typical query in computational linguistics could be *what tasks have CRFs been used for?*, which includes the elements of a typical schema for searching research papers; researchers want to find relationships between a technique and its applications (Gupta and Manning, 2011). Answers to this query can be found in various forms in published papers, for example,
(1) CRF-based POS tagging has achieved state-of-the-art accuracy.
(2) CRFs have been successfully applied to sequence labeling problems including POS tagging and named entity recognition.
(3) We apply feature reduction to CRFs and show its effectiveness in POS tagging.
(4) This study proposes a new method for the efficient training of CRFs. The proposed method is evaluated for POS tagging tasks.

Note that the same semantic relation, i.e., the use of CRFs for POS tagging, is expressed by various syntactic constructs: internal structures of the phrase in (1), clause-level structures in (2), inter-clause structures in (3), and discourse-level structures in (4). This implies that an integrated framework is required to represent semantic relations for phrase-level, clause-level, inter-clause level, and discourse-level structures. Another interesting fact is that we can recognize various fragments of information from single texts. For example, from sentence (1), we can identify *CRF is applied to POS tagging*, *state-of-the-art accuracy is achieved for POS tagging*, and *CRFs achieve high POS tagging accuracy*, all of which is valuable content for different search requests. This indicates that we need a framework that can cover (almost) all content in a text.

In this paper we describe a new annotation scheme for formalizing typical schemas for representing relations among concepts in research papers, such as techniques, resources, and effects. Our study aims to establish a framework for representing the semantics of research papers to help construct intelligent search systems. In particular, we focus on the formalization of typical schemas that we believe exemplify common query characteristics.

From the above observations, we have developed the following criteria for our proposed framework: use the same scheme for annotating contents in all levels of linguistic structures, annotate (almost) all contents presented in texts, and capture relations necessary for surveying research papers. We investigated 71 computer science abstracts (498 sentences) and defined an annotation

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 140–148,
Sofia, Bulgaria, August 8-9, 2013. ©2013 Association for Computational Linguistics

scheme comprising 16 types of semantic relations.

Computer science is particularly suitable for our purpose because it is primarily concerned with abstract concepts rather than concrete entities, which are typically the primary focus of empirical sciences such as physics and biology. In addition, computer and computational methods can be applied to an extraordinarily wide range of topics; computer science papers might discuss a bus timetable (for automatic optimization), a person’s palm (as a device for projecting images), or looking over another person sGshoulder (to obtain passwords). Therefore, to annotate all computer science papers, we cannot develop predefined entity ontologies, which is the typical approach taken in biomedical text mining (Kim et al., 2011).

However, most computer science papers have characteristic schemata: the papers describe a problem, postulate a method, apply the method to the problem using particular data or devices, and perform experiments to evaluate the method. The typical schemata clearly represent the structure of interests in this research field. Therefore, we can focus on typical schemata, such as *application of a method to a problem* and *evaluation of a method for a task*. As we will demonstrate in this paper, the proposed annotation scheme can cover almost all content, from phrase levels to discourse levels, in computer science papers.

Note that this does not necessarily mean that our framework can only be applied to computer science literature. The characteristics of the schemata described above are universal in contemporary science and engineering, and many other activities in human society. Thus, the framework presented in this study can be viewed as a starting point for research focusing on representative schemata of human activities.

2 Related Work

Traditionally, research on searching research papers has focused more on the social aspects of papers and their authors, such as citation links and co-authorship analysis implemented in the aforementioned services. Recently, research on content-based analysis of research papers has been emerging.

For example, methods of document zoning have been proposed for research papers in biomedicine (Mizuta et al., 2006; Agarwal and Yu, 2009; Liakata et al., 2010; Guo et al., 2011; Varga et al., 2012), and chemistry and computational linguistics (Teufel et al., 2009). Zoning provides a sentence-based information structure of papers to help identify the components such as the proposed method and the results obtained in the study. As such, zoning can narrow down the sections of a paper in which the answer to a query can be found. However, zoning alone cannot always capture the relation between the concepts described in the sections as it focuses on relation at a sentence level. For example, the examples (1), (2), (3) in the previous section require intra-sentence analysis to capture the relation between *CRF* and *POS tagging*. Our annotation scheme, which can be seen as conplementary to zoning, attempts to provide a structure for capturing the relationship between concepts at a finer-grained level than a sentence.

Establishing semantic relations among scientific papers has also been studied. For example, the ACL Anthology Searchbench (Schäfer et al., 2011) provides querying by predicate-argument relations. The system accepts specifications of subject, predicate, and object, and searches for texts that semantically match the query using the results from an HPSG parser. It can also search by topics automatically extracted from the papers. Gupta and Manning (2011) proposed a method for extracting *Focus*, *Domain*, and *Technique* from papers in the ACL anthology: *Focus* is a research article’s main contribution, *Domain* is an application domain, and *Technique* is a method or a tool used to achieve the *Focus*. The change in these aspects over time is traced to measure the influence of research communities on each other. Fukuda et al. (2012) developed a method of technical trend analysis that can be applied to both patent applications and academic papers, using the distribution of named entities. However, as processes and functions are key concepts in computer science, elements are often described in a unit with its own internal structures which include data, systems, and other entities as substructures. Thus, technical concepts such as technique cannot be captured fully by extracting named entities. Gupta and Manning (2011) analyzed the internal structures of concepts syntactically using a dependency parser, but did not further investigate the structure semantically.

In addition to the methodological aspects of research, i.e., what techniques are applied to what domain, a research paper can include other information that we also want to capture, such as how the author evaluates current systems and methods or the previous efforts of others. An attempt to identify the evaluation and other *meta*-aspects of scientific papers was made by Thompson et al. (2011), which, on top of the biomedical events annotated in the GENIA event corpus (Kim et al., 2008), annotated meta-knowledge such as the certainty level of the author, polarity (positive–negative), and manner (strong–weak) of events, as well as source (whether the event is attributed to the current study or previous studies), along with the clue mentioned in the text. For in-domain relations within and between the events, they relied on the underlying GENIA annotation, which maps events and their participants to a subset of Gene Ontology (The Gene Ontology Consortium, 2000), a standard ontology in genome science.

We cannot assume the existence of standard domain ontology in the variety of domains to which computer systems are applied, as was mentioned in Section 1. On the other hand, using domain-general linguistic frameworks, such as FrameNet (Ruppenhofer et al., 2006) or the Lexical Conceptual Structure (Jackendoff, 1990) is also not satisfactory for our purpose. These frameworks attempt to identify the relations lexicalized by verbs and their case arguments; however, they do not consider discourse or other levels of linguistic representation. In addition, relying on a linguistic theory requires that annotators understand linguistics. Most computer scientists, the best candidates for performing the annotation task, would not have the necessary knowledge of linguistics and would require training, which would increase costs for corpus annotation.

3 Annotation Scheme

The principle is to employ a uniform structure to represent semantic relations in scientific papers in phrase-level, clause-level, inter-clause level, and discourse-level structures. For this purpose, a bottom-up strategy that identifies relations between the entities mentioned is used. This strategy is similar to dependency parsing/annotation, which identifies the relations between constituents to find the overall structure of sentences.

We did not want the relations to be unconditionally concrete and domain-specific, because, as mentioned in the previous section, new concepts and relations that may not be expressed by predefined (concrete, domain-specific) concepts and relations may be created. For the same reason, we did not set specific entity types on the basis of domain ontology. We simply classified entities as “general object,” “specific object,” and “measurement.”

> In this paper, we propose a novel strategy for parallel preconditioning of large scale linear systems by means of a two-level approximate inverse technique with AISM method. According to the numerical results on an origin 2400 by using MPI, the proposed parallel technique of computing the approximate inverse makes the speedup of about 136.72 times with 16 processors.

Figure 1: Sample Abstract

To illustrate our scheme, consider the two-sentence abstract[1] shown in Figure 1[2].

In the first sentence, we can read that a method called *two-level approximate inverse* is used for parallel preconditioning (1), the preconditioning is applied to large-scale linear systems, the AISM method is a subcomponent or a substage of the two-level technique, and the author claims that the use of two-level approximate inverse is a novel strategy.

In the second sentence, we can read that the author has conducted a numerical experiment, the experiment was conducted on an origin 2400 (a computer system), message Passing Interface (MPI, a standardized method for message passing) was used in the experiment, the proposed parallel technique was 136.72 times quicker than existing methods, and the speedup was achieved using 16 processors.

In addition, by comparing the two sentences, we can determine that *the proposed parallel technique* in the second sentence refers to the parallel preconditioning using two-level approximate inverse mentioned in the first sentence. Consequently, we can infer the author’s claim that the parallel preconditioning using two-level approximate inverse achieved 136.72 times speedup.

We define binary relations including `APPLY_TO`(A, B) (A method A is applied to achieve the purpose B or used for doing B), `EVALUATE`(A, B) (A is evaluated as

[1] Linjie Zhang, Kentaro Moriya and Takashi Nodera. 2008. Two-level Parallel Computation for Approximate Inverse with AISM Method. IPSJ Journal, 48 (6): 2164-2168.

[2] Although the annotation was done for abstracts in Japanese, we present examples in English except where we discuss issues that we believe are specific to Japanese.

APPLY_TO(*two-level approximate inverse*, *parallel preconditioning*)

APPLY_TO(*parallel preconditioning*, *large scale linear systems*)

SUBCONCEPT(*AISM method*, *two-level approximate inverse*)

EVALUATE(*two-level approximate inverse*, *novel*)

RESULT(*numerical results*, *136.72 times speedup*)

CONDITION(*origin 2400*, *136.72 times speedup*)

APPLY_TO(*MPI*, *numerical results*)

EVALUATE(*the proposed parallel technique*, *136.72 times speedup*)

CONDITION(*16 processors*, *136.72 times speedup*)

EQUIVALENCE(*the proposed parallel technique*, *two-level approximate inverse*)

Figure 2: Relations Found in the Sentences in Figure 1

B), SUBCONCEPT(*A*, *B*) (*A* is a part of *B*), RESULT(*A*, *B*) (The result of experiment *A* is *B*), CONDITION(*A*, *B*) (The condition *A* holds in situation *B*), and EQUIVALENCE(*A*, *B*) (*A* and *B* refer to the same entity), with which we can express the relations mentioned in the example, as shown in Figure 2.

Note that it is *the use of two-level approximate inverse for parallel preconditioning*(*A*) that the author claims to be novel. However, the relation in *A* is already represented by the first APPLY_TO relation. Consequently, it is sufficient to annotate the EVALUATE relation between *two-level approximate inverse* and *novel*. This is approximately equivalent to paraphrasing *the use of two-level approximate inverse for parallel preconditioning is novel* as *two-level approximate inverse used for parallel preconditioning is novel*. The same holds for the equivalence relation involving *the proposed method*.

Expressing the content as the set of relations facilitates discovery of a concept that plays a particular role in the work. For example, if a reader wants to know the method for achieving parallel preconditioning, X, which satisfies the relation APPLY_TO(*X*, *parallel preconditioning*) must be searched for. By using the APPLY_TO relations mentioned in Figure 2 and inference on an *is-a* relation expressed by the SUBCONCEPT, we can obtain the result that *AISM method* is used for *parallel preconditioning*.

After a series of trial annotations on 71 abstracts from the IPSJ Journal (a monthly peer-reviewed journal published by the Information Processing Society of Japan), the following tag set was fixed. The annotation was conducted by the two of the authors of this paper.

3.1 Entity and Relation Types

The current tag set has 16 relation types and three entity types. An entity is whatever can be an argument or a participant in a relation. Entity types are OBJECT, MEASURE, or TERM, as shown in Table 1. Note that, unlike most schemes where the term *entity* refers to a nominal (named entity), in our scheme, almost all syntactic types of content words can be an entity, including numbers, verbs, adjectives, adverbs, and even some auxiliaries. The 16 types of relations are shown in Table 2. They are binary relations are directed from *A* to *B*.

Type	Definition	Example
OBJECT	the name of concrete entities such as a system, a person, and a company	Origin 2400, SGI
MEASURE	value, measurement, necessity, obligation, expectation, and possibility	novel, 136.72
TERM	any other	

Table 1: Entity Tags

All relations except EVALUATE COMPARE, and ATTRIBUTE can hold between any types of entity. EVALUATE and COMPARE relations hold between an entity (of any type) and an entity of the MEASURE type. The entities involved in an ATTRIBUTE relation must not be of the MEASURE type.

The INPUT and OUTPUT relations were introduced to deal with the distinction between the data and method used in computer systems. We extend the use of the scheme to annotate the inner structure of sentences and predicates, by establishing the relations between verbs and their case elements. For example, in *automatically generated test data*, obviously *test data* is an output of the action of *generate*, and *automatically* is the manner of generation. We annotate the *test data* as an OUTPUT and *automatically* as an ATTRIBUTE of *generate*. In another example, *a protocol that combines biometrics and zero-knowledge proof*, the protocol is the product of an action of combining biometrics and zero-

Type	Definition	Example
APPLY_TO(*A*, *B*)	A method *A* is applied to achieve the purpose *B* or used for conducting *B*	CRF_A-based $tagger_B$
RESULT(*A*, *B*)	*A* results in *B* in the sense that *B* is either an experimental result, a logical conclusion, or a side effect of *A*	$experiment_A$ shows the $increase_B$ in F-score compared to the baseline
PERFORM(*A*, *B*)	*A* is the agent of an intentional action *B*	a frustrated $player_A$ of a $game_B$
INPUT(*A*, *B*)	*A* is the input of a system or a process *B*, *A* is something obtained for *B*	$corpus_A$ for $training_B$
OUTPUT(*A*, *B*)	*A* is the output of a system or a process*B*, *A* is something generated from *B*	an $image_a$ $displayed_B$ on a palm
TARGET(*A*, *B*)	*A*is the target of an action *B*, which does not suffer alteration	to $drive_B$ a bus_A
ORIGIN(*A*, *B*)	*A* is the starting point of action *B*	to $drive_B$ from $Shinjuku_A$
DESTINATION(*A*, *B*)	*A* is the ending point of action *B*	an image $displayed_B$ on a $palm_A$
CONDITION(*A*, *B*)	The condition *A* holds in situation *B*, e.g, time, location, experimental condition	a $survey_B$ conducted in $India_a$
ATTRIBUTE(*A*, *B*)	*A* is an attribute or a characteristic of *B*	$accuracy_A$ of the $tagger_B$
STATE(*A*, *B*)	*A* is the sentiment of a person *B* other than the author, e.g. a user of a computer system or a player of a game	a $frustrated_A$ $player_B$ of a game

EVALUATE(*A*, *B*)
COMPARE(*C*, *B*)
	A is evaluated as *B* in comparison to *C*	experiment shows an $increase_B$
		in $F-score_A$ compared to the $baseline_C$

SUBCONCEPT(*A*, *B*)	*A* *is-a*, or is a *part-of* *B*	a $corpus_A$ such as PTB_a
EQUIVALENCE(*A*, *B*)	terms *A* and *B* refer to the same entity: definition, abbreviation, or coreference	DoS_B ($denial-of-service_A$) attack
SPLIT(*A*, *B*)	a term is split by parenthetical expressions into *A* and *B*	DoS_B (denial-of-service) $attack_A$

Table 2: Relation Tags

knowledge proof. Therefore, both *biometrics* and *zero-knowledge proof* are annotated as INPUTs of *combines*, and *protocol* is annotated as OUTPUT of *combines*. This scheme is not only used for computer-related verbs, but is further extended to any verb phrases or phrases with nominalized verbs. In *change in a situation*, *situation* is annotated as both INPUT and OUTPUT of *change*. It is as if we regard *change* as a machine that changes something, and when we input a situation, the *change-machine* processes it and output a different situation. Similarly, in *evolution of mobile phones*, *mobile phones* is annotated as both INPUT and OUTPUT of *evolution*. Here we regard *evolution* as a machine, and when we input (old-style) mobile phones, the *evolution-machine* processes them and outputs (new-style) mobile phones. We have found that a wide variety of predicates can be interpreted using these relations.

3.2 Other Features

Although we aim to annotate all possible relations mentioned, some conventions are introduced to reduce the workload.

First, we do not annotate the structure within entities. No nested entities are allowed, and compound words are treated as a single word. In addition, polarity (negation) is not expressed as a relation but as a part of an entity. We assume that the internal structure of entities can be analyzed by mechanisms such as technical term recognition. On the other hand, nested and crossed relations are allowed.

Second, we do not annotate words that indicate the existence of relations. This is because the relations are usually indicated by case markers and punctuation [3] and marking them up was found to be a considerable mental workload. In addition, words and phrases that directly represent the relations themselves are not annotated as entities. For example, in *CG iteration was applied to the problem*, we directly *CG relation* and *the problem* directly with APPLY_TO and skip the phrase *was applied to*.

Third, relations other than EQUIVALENCE and SUBCONCEPT are annotated within a sentence. We assume that the discourse-level relation can be inferred by the composition of relations.

In addition, the annotation of frequent verbs and their case elements was examined in the trial process. Verbs were classified, according to the pattern of the annotated relation with the case elements. For example, verbs semantically similar to *assemble* and *compile* form a class. The semantic role of the direct object of these verbs varies by context. For example, the materials in phrases like *compile source codes* or the product in phrases like

[3] This is in the case with Japanese. In languages such as English, there may be no trigger words, as the semantic relations are often expressed by the structure of sentences.

compile the driver from the source codes. In our scheme, the former is the `INPUT` of the verb, and the latter is the `OUTPUT` of the verb. Another example is the class of verbs that includes *learn* and *obtain*. The direct object (what is learned) is the `INPUT` to the system but is also the result or an output of the learning process. In such cases, we decided that both `INPUT` and `OUTPUT` should be annotated between the verb and its object.

Other details of annotation fixed in the process of trial annotation include:
1) The span of entities, which is determined to be the longest possible sequences delimited by case suffix (*-ga*,*-wo*, etc.) in the case of nominals and to separate the *-suru* suffix of verbs and the *-da* suffix of adjectives but retain other conjugation suffixes;
2) How to annotate evaluation sentences involving nouns derived from adjectives that imply evaluation and measurement, such as *necessity*, *difficulty*, and *length*. The initial agreement was that we would consider that they lose `MEASURE`-ness when nominalized; however, with the similarity of Japanese expressions *hitsuyou/mondai de aru* (is necessary/problematic) and *hitsuyou/mondai ga aru*(there is a necessity/problem), there was confusion about which word should be the `MEASURE` argument necessary for the `EVALUATE` relation. It was determined that, for example, in *hitsuyou/mondai de aru*, *de aru*, a copula, is ignored and *hitsuyou/mondai* is the `MEASURE`. In *hitsuyou/mondai ga aru*, *aru* is the `MEASURE`;
3) How to annotate phrases like *the tagger was better in precision*, where it can be understood that *the system* is evaluated as being *better in precision*. While what is actually measured in the evaluation process described in the paper is the precision (an attribute) of the tagger and the sentence has almost the same meaning as *the tagger's precision was better*, the surface (syntactic) subject of *is better* is *the tagger*. This can lead to two possibilities for the target of the `EVALUATE` relation. We decided that the `EVALUATE` relation holds between *precision* and *better*, and the `ATTRIBUTE` relation holds between *precision* and *tagger*, as illustrated in Figure 3.

A set of annotation guidelines was compiled as the result of the trial annotation, including the classifications and the pattern of annotation on frequent verbs and their arguments.

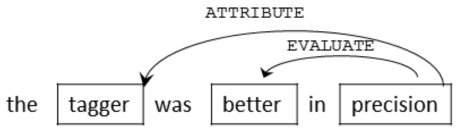


Figure 3: Annotation of *the tagger was better in precision*

Entity			Relation		
	Conunt	%		Conunt	%
Total	1895	100.0	Total	2269	100.0
OK	1658	87.5	OK	1110	48.9
Type	56	3.0	Type	250	11.0
Span	67	3.5	Direction	6	0.3
			Direction+Type	106	4.7
None	114	6.0	None	797	35.1

Table 3: Tag Counts

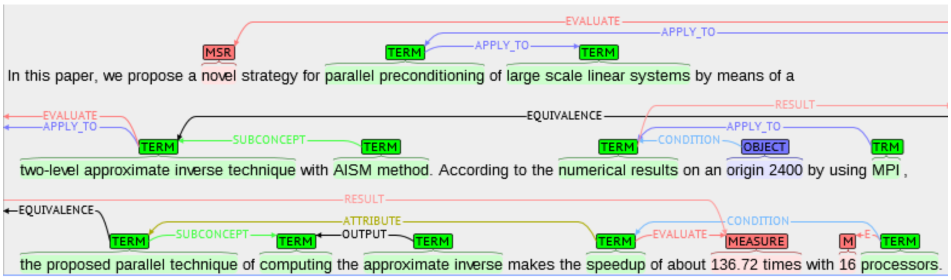
4 Annotation Experiment

We conducted an experiment on another 30 abstracts (197 sentences) from the IPSJ Journal. The two annotators who participated in the development of the guidelines annotated the abstracts independently, and inter-annotator discrepancy was checked. The annotation was performed manually using the brat annotation tool(Stenetorp et al., 2012). No automatic preprocessing was performed. Figure 4 shows the annotation results for the abstract shown in Figure 1. The 30 pairs of annotation results were aligned automatically; The results are shown in Tables 3, 4, and 5.

Table 3 shows the matches between the two annotators. "Total" denotes the count of entities/relations that at least one annotator found, "OK" denotes complete matches, "Type" denotes cases where two annotations on the same span have different entity/relation types, "Span" denotes entities where two annotations partially overlap, "Direction" denotes the count of relations where (only) the direction is different, and "Direction+Type"denotes relations where the same pair of entities were in different types of relation and in opposite directions, and "None" denotes cases where no counterpart was found in the other result.

Tables 4 and 5 are the confusion matrices for entity type and relation type, respectively. The differences in the span and direction are ignored. Agreement in F-score calculated in the same manner as in Brants (2000) for each relation is shown in column F, with the overall (micro-average) F-score shown in the bottom row of column F.

If we assume the number of cases that none of



In this paper, we propose a novel strategy for parallel preconditioning of large scale linear systems by means of a two-level approximate inverse technique with AISM method. According to the numerical results on an origin 2400 by using MPI , the proposed parallel technique of computing the approximate inverse makes the speedup of about 136.72 times with 16 processors.

Figure 4: Annotation Results with brat

	TERM	OBJECT	MEASURE	NONE	Total	F(%)
TERM	1458	2	38	14	1512	94.9
OBJECT	0	17	0	0	17	94.4
MEASURE	28	0	238	18	284	83.8
None	74	0	8	X	82	
Total	1560	19	284	32		93.0

Table 4: Confusion Matrix for Entity

the annotators recognized (the value of the cell X in the tables) to be zero, the observed agreement and Cohen's κ coefficient are 90.3% and 70.0% for entities, and 49.3% and 43.5% for relations, respectively. If we ignore the count for the cases where one annotator did not recognize the entity/relation ("None" rows and columns in the tables), the observed agreement and κ are 96.1% and 89.3% for entities, and 76.1% and 74.3% for relations, respectively. The latter statistics indicate the agreement on types for entities/relations that both annotators recognized.

These results show that entity annotation was consistent between the annotators but the agreement for relation annotation varied, depending on the relation type. Table 5 shows that agreement for DESTINATION, ORIGIN, EVALUATE, and SPLIT was reasonably high, but was low for CONDITION and TARGET. The rise in agreement (simple and κ) by excluding cases where only one annotator recognized the relation indicate that the problem is recognition, rather than classification, of relations[4].

From the investigation of the annotated text, the following was found:

(1) ATTRIBUTE/CONDITION decision was inconsistent in phrases involving EVALUATE relation, such as *the disk space is smaller for the image* (Figure 5). The EVALUATE relation between *the disk space* and *smaller* was agreed; however, the two annotators recognized different relations between *the image* and other words. One annotator recognized the ATTRIBUTE relation between *the disk space* and *the image* ("the disk space *as a feature of* the image is smaller"). The other recognized the CONDITION relation between *the image* and *smaller* ("the disk space is smaller *in the case of* the image").

(2) We were not in complete agreement about skipping phrases that directly represent a relation. The expressions to be skipped in the 71 trial abstracts were listed in the guidelines; however, it is difficult to exhaust all such expressions.

(3) In the case of some verbs, an argument can be INPUT and OUTPUT simultaneously (Section 3.1). We agreed that an object that undergoes alteration in a process should be tagged as both INPUT and OUTPUT but one that does not undergo alteration or which is just moved is the TARGET. Conflicts occurred for verbs that denote prevention of some situations such as *prevent*, *avoid*, and *suppress*, as illustrated in Figure 6. One annotator claimed that the possibility of DoS attacks is reduced to zero; hence the argument of the verb should be annotated with INPUT and OUTPUT. The other claims that since the DoS attack itself does not change, it is a TARGET.

(4) In a coordination expression, logical inference may be implicitly stated. For example, in *it requires the linguistic knowledge and is costly*, the reason for *costly* is likely to be the need for linguistic knowledge, i.e., employment of an expert linguist. However, the relation is not readily apparent. We wanted to capture the relation in such cases, but the disagreement shows that it is difficult to judge such a relation consistently.

(5) The decision on whether to split expressions like *XX dekiru* and *XX kanou* (can/able to *XX*) was also problematic. The guideline was to split them. This contradicts the decision for the compound words in general that we do not split them; however, we determined that *dekiru/kanou* cases had

[4] The same observation was true for entities

	APP	ATT	COMP	COND	DEST	EQU	EVAL	IN	ORIG	OUT	PER	RES	SPL	STA	SUB	TAR	None	Total	F(%)
APPLY_TO	136	9	0	2	1	1	2	10	1	0	0	3	0	0	1	0	65	231	53.0
ATTRIBUTE	14	154	0	19	6	0	9	5	1	0	7	1	0	0	3	0	28	247	59.7
COMPARE	0	0	6	1	0	0	0	0	0	0	0	0	0	0	0	0	4	11	54.5
CONDITION	4	11	1	77	0	0	1	4	0	0	0	5	0	0	0	0	49	152	48.7
DESTINATION	6	0	0	0	39	0	0	0	0	1	0	0	0	0	0	0	4	50	77.2
EQUIVALENCE	4	1	0	1	0	54	0	0	0	0	0	0	0	0	4	0	23	87	60.0
EVALUATE	0	11	0	0	0	0	215	3	0	9	0	0	0	0	0	1	41	280	76.1
INPUT	12	2	0	0	0	1	4	96	0	11	0	0	0	0	0	9	15	150	58.7
ORIGIN	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	2	18	78.0
OUTPUT	2	1	0	3	0	0	4	23	0	141	0	0	0	0	0	18	37	229	56.5
PERFORM	1	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	2	22	74.5
RESULT	8	1	0	0	0	0	1	1	0	0	0	38	0	0	0	0	22	71	54.3
SPLIT	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	80.0
STATE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SUBCONCEPT	14	10	0	3	0	4	5	0	0	2	0	0	0	0	81	0	34	153	58.1
TARGET	6	2	1	3	2	0	7	12	0	14	1	0	0	0	0	42	6	96	47.7
None	75	67	3	55	3	33	37	23	5	92	2	22	1	0	37	10	X	465	
Total	282	269	11	164	51	93	285	177	23	270	29	69	3	0	126	80	332		59.8

Table 5: Confusion Matrix for Relation

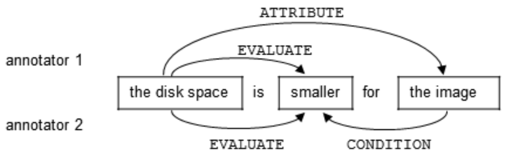


Figure 5: ATTRIBUTE/CONDITION Disagreement

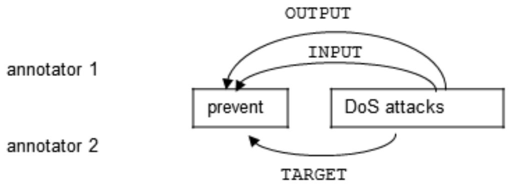


Figure 6: INPUT/OUTPUT/TARGET Disagreement

to be exceptions because the possibility of *XX* is expressed by *dekiru/kanou* and it seemed natural to relate *XX* and *dekiru/kanou* with EVALUATE. Unfortunately, confusion about splitting them remains.

5 Conclusions

We set up a scheme to annotate the content of research papers comprehensively. Sixteen semantic relations were defined, and guidelines for annotating semantic relations between concepts using the relations were established. The experimental results on 30 abstracts show that fairly good agreement was achieved, and that while entity- and relation-type determination can be performed consistently, determining whether a relation exists between particular pairs of entities remains problematic. We also found several discrepancy patterns that should be resolved and included in a future revision of the guidelines.

Traditionally, in semantic annotation of texts in the science/engineering domains, corpus creators focus on specific types of entities or events in which they are interested. On the other hand, we did not assume such specific types of entities or events, and we attempted to design a scheme that annotates more general relations in computer science/engineering domain.

Although the annotation is conducted for computer science abstracts in Japanese, we believe the scheme can be used for other languages, or for the broader science/engineering domains. The annotated corpus can provide data for constructing comprehensive semantic relation extraction systems. This would be challenging but worthwhile since such systems are in great demand. Such relation extraction systems will be the basis for content-based retrieval and other applications, including paraphrasing and translation.

The abstracts annotated in the course of the experiment have been cleaned up and are available on request. We are planning to increase the volume and make the corpus widely available.

In the future, we will assess machine-learning performance and incorporate the relation extraction mechanisms into search systems. Comparison of the annotated structure and the structures that can be given by existing semantic theories could be an interesting theoretical subject for future research.

Acknowledgments

This study was partially supported by the Japan Ministry of Education, Culture, Sports, Science and Technology Grant-in-Aid for Scientific Research (B) No. 22300031.

References

Shashank Agarwal and Hong Yu. 2009. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23):3174–3180.

Rafael E. Banchs, editor. 2012. *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. Association for Computational Linguistics.

Thorsten Brants. 2000. Inter-annotator agreement for a German newspaper corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation*.

Satoshi Fukuda, Hidetsugu Nanba, and Toshiyuki Takezawa. 2012. Extraction and visualization of technical trend information from research papers and patents. In *Proceedings of the 1st International Workshop on Mining Scientific Publications*.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 273–283.

Sonal Gupta and Christopher D Manning. 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th IJCNLP*.

Ray Jackendoff. 1990. *Semantic Structures*. The MIT Press.

Jin-Dong Kim, Tomoko Ohta, and Jun ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of bionlp shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1–6.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for conceptualisation and zoning of scientific papers. In *Proceedings of LREC 2010*.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, 75(6):468–487.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute.

Ulrich Schäfer, Bernd Kiefer, Christian Spurk, Jörg Steffen, and Rui Wang. 2011. The ACL anthology searchbench. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 7–13.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL*.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502.

The Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2011. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12.

Andrea Varga, Daniel Preotiuc-Pietro, and Fabio Ciravegna. 2012. Unsupervised document zone identification using probabilistic graphical models. In *Proceedings of LREC 2012*, pages 1610–1617.