Overview of the Protein Coreference task in BioNLP Shared Task 2011

Ngan Nguyen
University of Tokyo
Hongo 7-3-1, Bunkyoku, Tokyo
nltngan@is.s.u-tokyo.ac.jp

Jin-Dong Kim
Database Center for Life Science
Yayoi 2-11-16, Bunkyo-ku, Tokyo
jdkim@dbcls.rois.ac.jp

Jun'ichi Tsujii
Microsoft Research Asia
5 Dan Ling Street, Haiian District, Beijing
jtsujii@microsoft.com

Abstract

This paper summarizes the Protein Coreference Resolution task of BioNLP Shared Task 2011. After 7 weeks of system development period, the task received final submissions from 6 teams. Evaluation results show that state-of-the-art performance on the task can find 22.18% of protein coreferences with the precision of 73.26%. Analysis of the submissions shows that several types of anaphoric expressions including definite expressions, which occupies a significant part of the problem, have not yet been solved.

1 Introduction

While named entity recognition (NER) and relation or event extraction are regarded as standard tasks of information extraction (IE), coreference resolution (Ng, 2010; Bejan and Harabagiu, 2010) is more and more recognized as an important component of IE for a higher performance. Without coreference resolution, the performance of IE is often substantially limited due to an abundance of coreference structures in natural language text, i.e. information pieces written in text with involvement of a coreference structure are hard to be captured (Miwa et al., 2010). There have been several attempts for coreference resolution, particularly for newswire texts (Strassel et al., 2008; Chinchor, 1998). It is also one of the lessons from BioNLP Shared Task (BioNLP-ST, hereafter) 2009 that coreference structures in biomedical text substantially hinder the progress of fine-grained IE (Kim et al., 2009).

To address the problem of coreference resolution in molecular biology literature, the Protein Coreference (COREF) task is arranged in BioNLP-ST 2011 as a supporting task. While the task itself is not an IE task, it is expected to be a useful component in performing the main IE tasks more effectively. To establish a stable evaluation and to observe the effect of the results of the task to the main IE tasks, the COREF task particularly focuses on finding anaphoric protein references.

The benchmark data sets for developing and testing coreference resolution system were developed based on various manual annotations made to the Genia corpus (Ohta et al., 2002). After 7 weeks of system development phase, for which training and development data sets with coreference annotation were given, six teams submitted their prediction of coreferences for the test data. The best system according to our primary evaluation criteria is evaluated to find 22.18% of anaphoric protein references at the precision of 73.26%.

This paper presents overall explanation of the COREF task, which includes task definition (Section 2), data preparation (Section 4), evaluation methods (Section 5), results (Section 7), and thorough analyses (Section 8) to figure out what are remaining problems for coreference resolution in biomedical text.

2 Problem Definition

This section provides an explanation of the coreference resolution task in our focus, through examples.

Figure 1 shows an example text segmented into four sentences, S2 - S5, where anaphoric coreferences are illustrated with colored extends and arrows. In the figure, protein names are highlighted in purple, T4 - T10, and anaphoric protein references, e.g. pronouns and definite noun phrases, are highlighted in red, T27, T29, T30, T32, of which the an-

Proceedings of BioNLP Shared Task 2011 Workshop, pages 74–82,
Portland, Oregon, USA, 24 June, 2011.

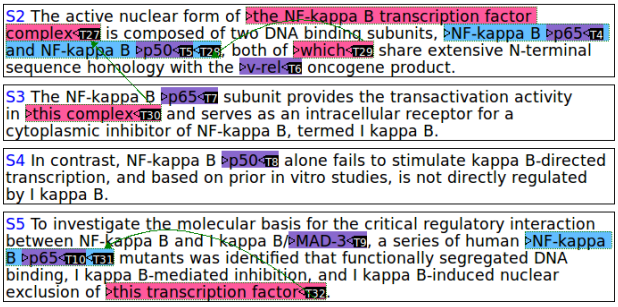

Figure 1: Protein coreference annotation

tecedents are indicated by arrows if found in the text. In the example, the definite noun phrase (NP), *this transcription factor* (T32), is a coreference to *p65* (T10). Without knowing the coreference structure, it becomes hard to capture the information written in the phrase, *nuclear exclusion of this transcription factor*, which is *localization of p65 (out of nucleus)* according to the framework of BioNLP-ST.

A standard approach would include a step to find candidate anaphoric expressions that may refer to proteins. In this task, pronouns, e.g. *it* or *they*, and definite NPs that may refer to proteins, e.g. *the transcription factor* or *the inhibitor* are regarded as candidates of anaphoric protein references. This step corresponds to *markable detection* and *anaphoricity determination* steps in the jargon of MUC. The next step would be to find the antecedents of the anaphoric expressions. This step corresponds to *anaphora resolution* in the jargon of MUC.

3 Task Setting

In the task, the training, development and test data sets are provided in three types of files: the text, the protein annotation, and the coreference annotation files. The *text* files contain plain texts which are target of annotation. The *protein annotation* files provide gold annotation for protein names in the texts, and the *coreference annotation* files provide gold annotation for anaphoric references to those protein names. The protein annotation files are given to the participants, together with all the training, development and test data sets. The coreference annotation files are not given with the test data set, and the task for the participants is to produce them automatically.

In protein annotation files, annotations for protein names are given in a stand-off style encoding. For example, those highlighted in purple in Figure 1 are protein names, which are given in protein annotation files as follows:

```
T4  Protein 275 278 p65
T5  Protein 294 297 p50
T6  Protein 367 372 v-rel
T7  Protein 406 409 p65
T8  Protein 597 600 p50
T9  Protein 843 848 MAD-3
T10 Protein 879 882 p65
```

The first line indicates *there is a protein reference in the span that begins at 275th character and ends before 278th character, of which the text is "p65", and the annotation is identified by the id, "T4"*

The coreference annotation files include three sort of annotations. First, annotations for anaphoric protein references are given. For example, those in red in Figure 1 are anaphoric protein references:

```
T27 Exp 179 222   the N.. 215 222 complex
T29 Exp 307 312   which
T30 Exp 459 471   this .. 464 471 complex
T32 Exp 1022 1047 this .. 1027 1047 tra..
```

The first line indicates that *there is an anaphoric protein reference in the specified span, of which the text is "the NF-kappa B transcription factor complex" (truncated due to limit of space)*, and that *its minimal expression is "complex"*. Second, noun phrases that are antecedents of the anaphoric references are also given in the coreference annotation file. For example, T28 and T31 (highlighted in blue) are antecedents of T29 and T32, respectively, and thus given in the file:

```
T28 Exp 264 297   NF-ka..
T31 Exp 868 882   NF-ka..
```

Third, the coreference relation between the anaphoric expressions and their antecedents are given in predicate-argument expressions[1]:

```
R1  Coref Ana:T29 Ant:T28 [T5, T4]
R2  Coref Ana:T30 Ant:T27
R3  Coref Ana:T32 Ant:T31 [T10]
```

The first line indicates *there is a coreference relation, R1, of which the anaphor is T29 and the antecedent is T28, and the relation involves two protein names, T5 and T4*.

Note that, sometimes, an anaphoric expression, e.g. *which* (T29), is connected to more than one protein names, e.g. *p65* (T4) and *p50* (T5). Sometimes, coreference structures do not involve any specific protein names, e.g. T30 and T27. In order

[1]Due to limitation of space, argument names are abbreviated, e.g. "Ana" for "Anaphora", and "Ant" for "Antecedent"

to establish a stable evaluation, our primary evaluation will focus only on coreference structures that involve specific protein names, e.g. T29 and T28, and T32 and T31. Among the three, only two, R1 and R3, involves specific protein references, T4 and T5, and T10. Thus, finding of R2 will be ignored in the primary evaluation. However, those not involving specific protein references are also provided in the training data to help system development, and will be considered in the secondary evaluation mode. See section 5 for more detail.

4 Data Preparation

The data sets for the COREF task are produced based on three resources: MedCO coreference annotation (Su et al., 2008), Genia event annotation (Kim et al., 2008), and Genia Treebank (Tateisi et al., 2005). Although the three have been developed independently from each other, they are annotations made to the same corpus, the Genia corpus (Kim et al., 2008). Since COREF was focused on finding anaphoric references to proteins (or genes), only relevant annotations were extracted from the MedCO corpus though the following process:

1. From MedCo annotation, coreference entities that were pronouns and definite base NPs were extracted, which became candidate anaphoric expressions. The base NPs were determined by consulting Genia Tree Bank.
2. Among the candidate anaphoric expressions, those that could not be protein references were filtered out. This process was done by checking the head noun of NPs. For example, definite NPs with "cell' as their head noun were filtered out. The remaining ones became candidate protein coreferences.
3. The candidate protein coreferences and their antecedents according to MedCo annotation were included in the data files for COREF task.
4. The protein name annotations from Genia event annotation were added to the data files to determine which coreference expressions involve protein name references.

Table 1 summarizes the coreference entities in the training, development, and test sets for COREF task. In the table, the anaphoric entities are classified into four types as follows:

RELAT indicates relative pronouns or relative adjectives, e.g. *that*, *which*, or *whose*.

PRON indicates pronouns, e.g. *it*.

Type		Train	Dev	Test
Anaphora	RELAT	1193	254	349
	PRON	738	149	269
	DNP	296	58	91
	APPOS	9	1	3
	N/C	11	1	2
Antecedent		2116	451	674
TOTAL		4363	914	1388

Table 1: Statistics of coreference entities in COREF data sets: N/C = not-classified.

DNP indicates definite NPs or demonstrative NPs, e.g. NPs that begin with *the*, *this*, etc.

APPOS indicates coreferences in apposition.

5 Evaluation

The coreference resolution performance is evaluated in two modes.

The *Surface coreference mode* evaluates the performance of finding anaphoric protein references and their antecedents, regardless whether the antecedents actually embed protein names or not. In other words, it evaluates the ability to predict the coreference relations as provided in the gold coreference annotation file, which we call *surface coreference links*.

The *protein coreference mode* evaluates the performance of finding anaphoric protein references with their links to actual protein names (*protein coreference links*). In the implementation of the evaluation, the chain of surface coreference linkes is traced until an antecedent embedding a protein name is found. If a protein-name-embedding antecedent is connected to an anaphora through only one surfs link, we call the antecedent a *direct protein antecedent*. If a protein-name-embedding anteceden is connected to an anaphora through more than one surface link, we call it an *indirect protein antecedent*, and the antecedents in the middle of the chain *intermediate antecedents*. The performance evaluated in this mode may be directly connected to the potential performance in main IE tasks: the more the (anaphoric) protein references are found, the more the protein-related events may be found. For this reason, the protein coreference mode is chosen as the primary evaluation mode.

Evaluation results for both evaluation modes are

given in traditional precision, recall and f-score, which are similar to (Baldwin, 1997).

5.1 Surface coreference

A response expression is matched with a gold expression following partial match criterion. In particular, a response expression is considered correct when it covers the minimal boundary, and is included in the maximal boundary of expression. Maximal boundary is the span of expression annotation, and minimal boundary is the head of expression, as defined in MUC annotation schemes (Chinchor, 1998). A response link is correct when its two argument expressions are correctly matched with those of a gold link.

5.2 Protein coreference

This is the primary evaluation perspective of the protein coreference task. In this mode, we ignore coreference links that do not reference to proteins. Intermediate antecedents are also ignored.

Protein coreference links are generated from the surface coreference links. A protein coreference link is composed of an anaphoric expression and a protein reference that appears in its direct or indirect antecedent. Below is an example.

```
Example:
R1  Coref Ana:T29 Ant:T28 [T5, T4]
R2  Coref Ana:T30 Ant:T27
R3  Coref Ana:T32 Ant:T31 [T10]
R4  Coref Ana:T33 Ant:T32
```

In this example, supposing that there are four surface links in the coreference annotation file (T29,T28), (T30,T27), (T32,T31), and (T33, T32), in which T28 contains two protein mentions T5, T4, and T31 contains one protein mention T10; thus, the protein coreference links generated from these surface links are (T29,T4), (T29,T5), (T32,T10), and (T33, T10). Notice that T33 is connected with T10 through the intermediate expression T32.

Response expressions and generated response result links are matched with gold expressions and links correspondingly in a way similar to the surface coreference evaluation mode.

6 Participation

We received submissions from six teams. Each team was requested to submit a brief description of their team, which was summarized in Table 2.

Team	Member	Approach & Tools
UU	1 NLP	ML (Yamcha SVM, Reconcile)
UZ	5 NLP	RB (-)
CU	2 NLP	RB (-)
UT	1 biochemist	ML (SVM-Light)
US	2 AI	ML (SVM-Light)
UC	3 NLP, 1 BioNLP	ML (Weka SVM)

Table 2: Participation. UU = UofU, UZ = UZH, CU=ConcordU, UT = UTurku, UZ = UZH, US = Uszeged, UC = UCD_SCI, RB = Rule-based, ML = Machine learning-based.

TEAM	RESP	C	P	R	F
UU	86	63	73.26	22.18	34.05
UZ	110	61	55.45	21.48	30.96
CU	87	55	63.22	19.37	29.65
UT	61	41	67.21	14.44	23.77
US	259	9	3.47	3.17	3.31
UC	794	2	0.25	0.70	0.37

Table 3: Protein coreference results. Total number of gold link = 284. RESP=response, C=correct, P=precision, R=recall, F=fscore

The *tool* column shows the external tools used in resolution processing. Among these tools, there is only one team used an external coreference resolution framework, *Reconcile*, which achieved the state-of-the-art performance for supervised learning-based coreference resolution (Stoyanov et al., 2010b).

7 Results

7.1 Protein coreference results

Evaluation results in the protein coreference mode are shown in Table 3. The UU team got the highest f-score 34.05%. The UZ and CU teams are the second- and third-best teams with 30.96% and 29.65% f-score correspondingly, which are comparable to each other. Unfortunately, two teams, US and UC could not produce meaningful results, and the other four teams show performance optimized for high precision. It was expected that the 22.18% of protein coreferences may contribute to improve the performance on main task, which was not observed this time, unfortunately.

The first ranked system by UU utilized Recon-

TEAM	RESP	C	P	R	F
UU	360	43	11.94	20.48	15.09
UZ	736	51	6.93	24.29	10.78
CU	365	36	9.86	17.14	12.52
UT	452	50	11.06	23.81	15.11
US	259	4	1.54	1.90	1.71
UC	797	1	0.13	0.48	0.20

Table 4: Surface coreference results. Total number of gold link = 210. RESP=response, C=correct, P=precision, R=recall, F=fscore

	UU	UT
S-correct & P-missing	8	29
S-missing & P-correct	16	5

Table 5: Count of anaphors that have different status in different evaluation modes. S = surface coreference evaluation mode, P = protein coreference evaluation mode

cile which was originally developed for newswire domain. It supports the hypothesis that machine learning-based coreference resolution tool trained on different domains can be helpful for the bio medical domain; however, it still requires some adaptations.

7.2 Surface coreference results

Table 4 shows the evaluation results in the surface link mode. The overall performances of all the systems are low, in which recalls are much higher than the precisions. One possible reason of the low results is because most of the teams focus on resolving pronominal coreference; however, they failed to solve some difficult types of pronoun such as "it", "its", "these", "them", and "which", which occupy the majority of anaphoric pronominal expressions (Table 1). Definite anaphoric expressions were ignored by almost all of the systems (except one submission).

The results show that the protein coreference resolution is not a trivial task; and many parts remains challenging. In next section, we analyze about potential reason of the low results, and discuss possible directions for further improvement.

Ex 1	GOLD
T5	DQalpha and DQbeta trans heterodimeric HLA-DQ molecules
T6	such *trans-dimers*
T7	which
R1	T6 T5 [T3, T4]
R2	T7 T6
	RESP
T5	such trans-dimers
T6	which
R1	T6 T5
Ex 2	GOLD
T18	Five *members* of this family (MYC, SCL, TAL-2, LYL-1 and E2A)
T20	their
R3	T20 T18 [T3, T2, T5, T4]
	RESP
T19	Five members
T20	their
R2	T20 T19

Table 6: Example of surface-correct & protein-missing cases. Protein names are underlined, and the min-values are in italic.

8 Analysis

8.1 Why the rankings based on the two evaluation methods are not the same?

Comparing with the protein coreference mode, we can see the rankings based on two evaluation methods are different. In order to find out what led to this interesting difference, we further analyzed the submissions from the two teams UT and UU. The UT team achieved the highest f-score in the surface evaluation mode, but was in the fourth rank in the protein evaluation mode. Meanwhile, the score of UU team was slightly less than the UT team in the former mode, but got the highest in the later (Table 3 and Table 4). In other words, there is no clear correlation between the two evaluation results.

Because the two precisions in surface evaluation mode are not much different, the recalls were the main contribution in the difference of f-score. Analyzing the correct and missing examples in both evaluation modes, we found that there are anaphors whose surface links are correct, while the protein links with the same anaphors are evaluated as missing; and vice versa with missing surface links and correct protein links. Counts of anaphors of each

type are shown in Table 5. In this table, the cell at column *UT* and row *S-correct and P-missing* can be interpreted as following. There are 29 anaphors in the UT response whose surface links are correct but protein links are missing, which contributes positively to the recall in *surface coreference mode*, and negatively to that in *protein coreference mode*.

Table 6 shows two examples of *S-correct and P-missing*. In the first example, we can see that the gold antecedent proteins are contained in an indirect antecedent. Therefore, when the intermediate antecedent is correctly detected by the surface link *R1*, but the indirect antecedent is not detected, the anaphor is not linked to it antecedent proteins "DQalpha" and "DQbeta". Another reason is because response antecedents do not include antecedent proteins. This is actually the problem of expression boundary detection. An example of this is example 2 (Table 6), in which the response surface link *R2* is correct, but the protein links to the four proteins are not detected, because the response antecedent "five members" does not include the protein mentions "SCL, TAL-2, LYL-1 and E2A". However, the response antecedent expression is correct because it contains the minimal boundary "members".

For *S-missing and P-correct*, we found that anaphors are normally directly linked to antecedent proteins. In other words, expression boundary is same as protein boundary. Another case is that response antecedents contain the antecedent proteins, but are evaluated as incorrect because the expression boundary of the response expression is larger than the gold expression. An example is shown in Table 7 where the response expression "a second GCR, termed GCRbeta" includes the gold expression "GCRbeta". Therefore, although the surface link is incorrect because the response expression is evaluated as incorrect, the protein coreference link receives a full score .

The difference reflects the characteristics of the two evaluation methods. The analysis result also shows the affect of markable detection or expression detection on the resolution evaluation result.

8.2 Protein coreference analysis

We want to see how well each system performs on each type of anaphor. However, the type information is not explicitly included in the response, so it has to be induced automatically. We done this by finding the first word of anaphoric expression; then, we combine it with *1* if the expression is a single-word expression, or *2* if the expression is multi-word, to create a sub type value for each anaphor of both gold and response anaphors. After that, subtypes are mapped with the anaphor types specified in Section 4 using the mapping in Table 10.

Ex 3	GOLD
T17	GCRbeta
T18	which
R2	T18 T17 [T4]
	RESP
T16	a second GCR, termed GCRbeta
T19	which
R2	T19 T16

Table 7: Examples of S-missing and P-correct

Protein coreference resolution results by sub type are given in Table 9 and 8. It can be easily seen in Table 9 which team performed well on which type of anaphor. In particular, the CU system was good at resolving the RELAT, APPOS and other types. The UU team performed well on the DNP type. And for the PRON type, UZ was the best team. In theory, knowing this, we can combine strengths of the teams to tackle all the types.

We analyzed false positive protein anaphora links to see what types of anaphora are solved by each system. The recalls in Table 11 are calculated based on the anaphor type information manually annotated in the gold data. Comparing with those in Table 9, there is a small difference due to the automatic induction of anaphoric types based on sub types. It can be seen in the table 11 that only 77.5 percent of RELAT-typed anaphora links were resolved (by CU team), although this type is supposed to be the easiest type. Examining the output data, we found that the system tends to choose the nearest expression as the antecedent of a relative pronoun; however, this is not always correct, as in the following examples from the UofU submission: "We also identified *functional Aiolos-binding sites*$_1a$ in the *Bcl-2 promoter*$_1b$, *which*$_1$ are able to activate the luciferase reporter gene.", and "Furthermore, the analysis of IkappaBalpha turnover demonstrated *an increased*

	PRON both-2	P- it-1	P- its-1	P- one-2	P- that-1	P- their-1	P- these-2	DNP this-2	D- those-1	RELAT which-1	R- whose-1	N/C
UU			36.4		64.4		2	13.3	18.2	62	5	30.8
UZ		46.2	35.7		53.3	7.1		12.5	5.4	59	66.7	15.4
CU					62					70.9	5	42.1
UT		9.5	36.8	10	34.6				9.5	5		30.8
US			13.9			22.9						
UC	28.6	9.1										

Table 8: Fine-grained results (f-score, %)

Team	PRON P	P- R	P- F	DNP P	D- R	D- F	RELAT P	R- R	R- F	Others P	O- R	O- F
UU	79.0	11.5	20.1	66.7	5.9	10.8	71.3	56.0	62.7	100.0	18.3	30.8
UZ	62.9	16.9	26.7	12.5	4.4	6.5	71.4	46.7	56.5	50.0	9.1	15.4
CU	_	_	_	_	_	_	64.6	68.0	66.2	50.0	36.4	42.1
UT	72.7	12.3	21.1	14.3	1.5	2.7	73.3	29.3	41.9	100.0	18.2	30.8
US	27.3	6.9	11.0	_	_	_	_	_	_	_	_	_
UC	9.1	1.5	2.6	_	_	_	_	_	_	_	_	_

Table 9: Protein coreference results by coreference type (fscore, %). P = precision, R = recall, F = f-score. O = Others.

TEAM	A	R	D	P	O
UU	0.0	62.0	5.7	11.1	0.0
UZ	0.0	49.3	4.3	17.0	0.0
CU	0.0	77.5	0.0	0.0	0.0
UT	0.0	32.4	1.4	11.9	14.3
US	0.0	0.0	0.0	6.7	0.0
UC	0.0	0.0	1.4	0.7	0.0

Table 11: Exact recalls by anaphor type, based on manual *type* annotation. A=APPOS, R=RELAT, D=DNP, P=PRON, O=OTHER

degradation of $\underline{IkappaBalpha}_2a$ in *HIV-1-infected* $cells_2b$ $that_2$ may account for the constitutive DNA binding activity.". Expressions with the same index are coreferential expressions. The *a* subscript indicates correct antecedent, and *b* subscript indicates the wrong one. In these examples, the relative pronoun *that* and *which* are incorrectly linked with the nearest expression, which is actually part of post-modifier or the correct antecedent expression.

For the DNP type, recall of the best system is less than 6 percent (Table 11), although it is an important type which occupies almost one fifth of all protein links (Table 1). There is only one team, the UC team, attempted to tackle *the* anaphor; however, it resulted in many spurious links. The other teams did not make any prediction on this type. A possible reason of this is because there are much more non-anaphoric definite noun phrases than anaphoric ones, which making it difficult to train an effective classier for anaphoricity determination. We have to seek for a better method for solving the DNP links, in order to significantly improve protein coreference resolution system.

Concerning the PRON type, Table 8 shows that except for *that-1*, no other figures are higher than 50 percent f-score. This is an interesting observation because pronominal anaphora problem has been reported with much higher results on other domains(Raghunathan et al., 2010), and also on other bio data (hsiang Lin and Liang, 2004). One of the reasons for the low recall is because target anaphoric pronouns in the bio domain are neutral-gender and third-person pronouns(Nguyen and Kim, 2008), which are difficult to resolve than other types of pronouns(Stoyanov et al., 2010a).

8.3 Protein coreference analysis - Intermediate antecedent

As mentioned in the task setting, anaphors can directly link to their antecedent, or indirectly link via one or more intermediate antecedents. We counted the numbers of correct direct and indirect protein coreference links in each submission (Table 12).

Sub type	Type	Count	Sub type	Type	Count	Sub type	Type	Count
both_1	PRON	2	both_2	PRON	4	either_1	PRON	0
it_1	PRON	17	its_1	PRON	61	one_2	PRON	1
such_2	DNP	2	that_1	RELAT	37	the_2	DNP	20
their_1	PRON	27	them_1	PRON	1	these_1	PRON	1
these_2	DNP	26	they_1	PRON	5	this_1	PRON	1
this_2	DNP	20	those_1	PRON	9	which_1	RELAT	37
whose_1	RELAT	1	whose_2	RELAT	0	(others)	N/C	11

Table 10: Mapping from sub type to coreference type. Count = number of anaphors

TEAM	A	R	R	D	D	P	P	O
	Di	Di	In	Di	In	Di	In	Di
UU		44		4		15		
UZ		35		2	1	23		
CU		54	1					
UT		22	1	1		16		1
US						8	1	
UC					1	1		
Total	1	64	7	65	5	126	9	7

Table 12: Numbers of correct protein coreference links by anaphor type and by number of antecedents, based on manual *type* annotation. A=APPOS, R=RELAT, D=DNP, P=PRON, O=Others. Di=direct, In=indirect.

APPOS and Others types do not have any intermediate antecedent, thus there is only one column marked with *D* (direct protein coreference link). We can see in this table that very few indirect links were detected. Therefore, there is place to improve our resolution system by focusing on detection of such links.

8.4 Surface coreference results

Because inclusion of all expressions was not a requirement of shared task submission, the submitted results may not contain expressions that do not involve in any coreference links. Therefore, it is unfair to evaluate expression detection based on the response expressions.

Evaluation results for anaphoricity determination are shown in Table 13. The calculation is performed as following. Supposing that every anaphor has a response link, the number of anaphors is number of distinct anaphoric expressions inferred from the response links, which is given in the first column. The total number of gold anaphors are also calculated in similar way. Since response expressions are lined with gold expressions before evaluation,

Team	Resp	Align	P	R	F
UU	360	94.2	19.4	33.3	24.6
UZ	736	75.8	22.0	77.1	34.2
CU	365	89.6	15.3	26.7	19.5
UT	452	92.0	18.1	39.0	24.8
US	259	9.3	6.2	7.6	6.8
UC	797	6.8	1.1	4.3	1.8

Table 13: Anaphoricity determination results. Total number of gold anaphors = 210. Resp = number of response anchors, Align = alignment rate(%), P = precision (%), R = recall (%), F = f-score (%)

we provided the alignment rate for reference in the second column of the table. The third and forth columns show the precisions and recalls. In theory, low anaphoricity determination precision results in many spurious response links, while low recall becomes the bottle neck for the overall coreference resolution recall. Therefore, we can conclude that the low performance of anaphoricity determination contribute to the low coreference evaluation results (Table 4, Table 3).

9 Conclusion

The coreference resolution supporting task of BioNLP Shared Task 2011 has drawn attention from researchers of different interests. Although the overall results are not good enough to be helpful for the main shared tasks as expected, the analysis results in this paper shows the coreference types which have and have not yet been successfully solved. Tackling the remained problems in expression boundary detection, anaphoricity determination and resolution algorithms for difficult types of anaphors such as definite noun phrases should be the future work. Then, it would be interesting to see how much coreference can contribute to event extraction.

References

B. Baldwin. 1997. Cogniac: High precision with limited knowledge and linguistic resources. In *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, pages 38–45, Madrid, Spain.

Cosmin Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422, Uppsala, Sweden, July. Association for Computational Linguistics.

Nancy Chinchor. 1998. Overview of MUC-7/MET-2. In *Message Understanding Conference (MUC-7) Proceedings*.

Yu hsiang Lin and Tyne Liang. 2004. Pronominal and sortal anaphora resolution for biomedical literature. In *In Proceedings of ROCLING XVI: Conference on Computational Linguistics and Speech Processing*.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from lterature. *BMC Bioinformatics*, 9(1):10.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*, pages 1–9.

Makoto Miwa, Rune Sætre, Jin-Dong Kim, and Jun'ichi Tsujii. 2010. Event extraction with complex event classification using rich features. *Journal of Bioinformatics and Computational Biology (JBCB)*, 8(1):131–146, February.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the ACL*, pages 1396–1411.

Ngan Nguyen and Jin-Dong Kim. 2008. Exploring domain differences for the design of a pronoun resolution system for biomedical texts. In *Proceedings of 22nd International Conference on Computational Linguistics (COLING-2008)*.

T Ohta, Y Tateisi, H Mima, and J Tsujii. 2002. Genia corpus: an annotated research abstract corpus in molecular biology domain. *Proceedings of the Human Language Technology Conference (HLT 2002), San Diego, California*, pages 73–77.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nate Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, October.

V. Stoyanov, C. Cardie, N. Gilbert, E. Riloff, D. Buttler, and D. Hysom. 2010a. Coreference resolution with reconcile. In *Proceedings of the Conference of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*.

V. Stoyanov, C. Cardie, N. Gilbert, E. Riloff, D. Buttler, and D. Hysom. 2010b. Reconcile: A coreference resolution platform. In *Tech Report - Cornell University*.

Stephanie Strassel, Mark Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda. 2008. Linguistic Resources and Evaluation Techniques for Evaluation of Cross-Document Automatic Content Extraction. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Jian Su, Xiaofeng Yang, Huaqing Hong, Yuka Tateisi, and Jun'ichi Tsujii. 2008. Coreference Resolution in Biomedical Texts: a Machine Learning Approach. In *Ontologies and Text Mining for Life Sciences'08*.

Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun'ichi Tsujii. 2005. Syntax annotation for the genia corpus. In *International Joint Conference on Natural Language Processing*, pages 222–227, Jeju Island, Korea, October.