Zone Identification in Biology Articles as a Basis for Information Extraction

Yoko MIZUTA and Nigel COLLIER
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo, Japan, 101-8430
{ymizuta, collier}@nii.ac.jp

Abstract

Information extraction (IE) in the biomedical domain is now regarded as an essential technique for the dynamic management of factual information contained in archived journal articles and abstract collections. We aim to provide a technique serving as a basis for pin-pointing and organizing factual information related to experimental results. In this paper, we enhance the idea proposed in (Mizuta and Collier, 2004); annotating articles in terms of rhetorical zones with shallow nesting. We give a qualitative analysis of the zone identification (ZI) process in biology articles. Specifically, we illustrate the linguistic and other features of each zone based on our investigation of articles selected from four major online journals. We also discuss controversial cases and nested zones, and ZI using multiple features. In doing so, we provide a stronger theoretical and practical support for our framework toward automatic ZI.

1 Introduction

Information extraction (IE) in the biomedical domain is now regarded as an essential technique for utilizing information contained in archived journal articles and abstract collections such as MEDLINE. Major domain databases often contain incomplete and inconsistent results. Also, a majority of the reported experimental results are only available in unstructured full-text format. These being combined, scientists need to check with source journal articles to obtain and confirm factual information. Furthermore, they often need to start with document retrieval and face an overwhelming number of candidate articles. Thus, the significance of dynamic management of factual information, specifically an integration and update of experimental results, is self-evident. It would not only save researchers much time used for retrieval and redundant experiments but also help them use the information more effectively. Given the limitations of manual work in terms of both efficiency and accuracy, IE focusing on factual information is of critical importance.

Researches in bioNLP have made major progress mainly in the extraction of bio-named entity biological interactions (e.g. Craven et al. 1999; Humphreys et al., 2000; Tanabe et al., 2002). But further progress aimed at pin-pointing and organizing factual information remains a challenge.

We aim to provide a basis for this purpose. As the first step, we proposed in (Mizuta and Collier, 2004) annotating biology texts in terms of rhetorical zones with a shallow nesting, and provided an annotation scheme. In this paper, we explore a qualitative analysis of zone identification (ZI) in biology articles and provide stronger support for our framework toward automatic annotation of zones. Specifically we; 1) illustrate the linguistic and other features of each zone, which have been extracted through our pilot study of a total of 20 articles randomly selected from four major online journals (EMBO, PNAS, NAR and JCB), 2) discuss controversial cases for ZI and nested annotation to elaborate the scheme, 3) discuss multiple features relevant to ZI, and 4) summarize the investigation and outline future steps related to machine learning and applications.

Previous work on rhetorical analysis of scientific articles focus on either; 1) hierarchical discourse relations between sentences (e.g. Mann and Thompson, 1987), 2) genre analysis within a descriptive framework (e.g. Swales 1990), or 3) ZI in a flat structure and a statistical evaluation of the annotation scheme from a machine learning perspective (e.g. Teufel and Moens, 2002). We follow the lines of (Teufel and Moens, 2002) and apply ZI to the domain of biology. But our approach is unique in that we focus on experimental results and on a qualitative analysis of ZI as a basis for automatic ZI.

2 Overview of the framework

2.1 The need for zone identification (ZI)

We discuss below the critical issues in bioNLP involved in pin-pointing and organizing factual information and show how ZI can be applied.

First, articles provide information in various rhetorical statuses (e.g. new vs. old results; own vs. previous work). Current IE relies on surface lexical

and syntactic patterns, neglecting the rhetorical status of information. Thus, we are in danger of extracting old results mixed with new ones.

(1) **Recent data suggest that** … ~ is involved in DPC removal in mammalian cells (ref.), … …**The data presented here suggest that** …

The data (1) provide statements in different rhetorical statuses (boldfaced by us). Preprocessing the text in terms of such information helps filter out old results (i.e. the first statement).

Secondly, so far the scope of bioNLP largely bear on abstracts. But arguably, the final goal should be full texts, given their much richer sources of information and the increasing ease of access (e.g. open access to collections such as PUBMED-central; online journals such as EMBO, PNAS, and JCB). This involves exploring new techniques because there are some essential differences from abstracts. Among others, full texts present much more complexity in the sentence structure and vocabulary (e.g. inserted phrases, embedded sentences, nominalization of verbs, more anaphoric expressions). Thus, we expect that the analysis of the whole text requires a much more complex set of patterns and algorithms, [1] resulting in errors. A solution to this problem is to identify the subset of the article relevant to further analysis at issue. For example, in order to extract certain kinds of biological interactions found by the author, we could skip statements about previous work as seen in the Introduction section.

Thirdly, experimental results make sense in their relation to the experimental goal and procedure. Also, there are usually a sequence of experiments performed, each of which obtains complex results. Therefore, it is important to extract a set of experimental results in an organized manner. This also helps identify the reference of demonstratives (e.g. *this*) and pronouns (e.g. *it*).

From these points of view, ZI in articles plays an essential role in extracting factual information of different sorts from different zone classes.

2.2 Characteristics of the framework

The idea underlying ZI in our sense contrasts with other, discourse relations-based notions (e.g. Mann et al. 1987; Kando 1999; van Dijk, 1980); we focus on the global type of information. For example, in our ZI, reference to previous work as background information remains as such whether it is supported or refuted by the author later in the article, whereas this difference plays an essential role in discourse relations-based analyses.

[1] A. Koike (at AVIRG 2004) reported that to extract the interactions between two biological elements from PUBMED abstracts, about 400 patterns were necessary.

The larger picture we have consists of 2 levels; 1) ZI, and 2a) analysis of zone interactions (e.g. discourse relations), or 2b) analysis on specific zones (i.e. extraction of biological interactions). In this paper we focus on the first step.

2.3 Annotation scheme

Our annotation scheme is proposed in (Mizuta et al., 2004), based on Teufel et al.'s (2002) scheme. Three major modifications are made; 1) a fine-grained OWN class based on the model of an experimental procedure which we identified across journals, 2) CNN and DFF classes to cover the relations between data/findings, and 3) nested annotation. The set of zone classes is as follows:

- BKG (Background): given information (reference to previous work or a generally accepted fact)
- PBM (Problem-setting): the problem to be solved; the goal of the present work/paper.
- OTL (Outline): a characterization/ summary of the content of the paper.
- TXT (Textual): section organization of the paper (e.g. "Section 3 describes our method").
- OWN: the author's own work:
  - ◊MTH (Method): experimental procedure;
  - ◊RSL (Result): the results of the experiment;
  - ◊INS (Insight): the author's insights and findings obtained from experimental results (including the interpretation) or from previous work
  - ◊IMP (Implication): the implications of experimental results (e.g. conjectures, assessment, applications, future work) or those of previous work
  - ◊ELS (Else): anything else within OWN.
- CNN (Connection): correlation or consistency between data and/or findings.
- DFF (Difference): a contrast or inconsistency between data and/or findings.

The basic annotation unit is a sentence, but in some cases it may be a phrase. In light of those cases which fit into multiple zones, we employ 2-level annotation. Empirical analysis indicates that even though zone classes are conceptually non-overlapping, an annotation unit may fit into multiple classes. That is, a linguistic unit (e.g. a sentence) may well represent complex concepts. Therefore, we consider that nested annotation is necessary, even though it complicates annotation.

3 Zone identification -1: Main features of each zone

Based on our sample annotation of full texts, we discuss the major features extracted from the data

of each zone class. Complex cases and the location of zones will be discussed in later sections.

3.1 BACKGROUND (BKG)

(1) In cells, DNA **is** tightly **associated** with …
(2) Ref. **suggested**/ **suggests** that ~
(3) A wide variety of restriction-modification (R-M) systems **have been discovered** ….

BKG has three tense variations; 1) simple present for a generic statement about background information (e.g. biological facts; reference to previous work), 2) simple past, and 3) present perfect, to mention the current relevance of previous work. A wider range of verbs are used to cover both biological and bibliographical facts.

Citations in the sentence-final position having as its scope the whole sentence signal BKG, but inter-sentential citations having a smaller scope do not.

3.2 PROBLEM SETTING (PBM)

There are two types of PBM.

(2) X has **not** been established/addressed
there has been **no** study on X
little is currently known about ~
there are very **limited** data concerning X
X **remain unclear**

The first type as illustrated above is observed in the I-section[2]; it addresses the problem to solve. It has a 'negative polarity' in that it mentions something missing in the current situation (e.g. knowledge, study, a research question). It contains vocabulary expressing negation or incompleteness (boldfaced). Tense variation is either simple present or present perfect, depending on the temporal interval referred to. The range of verbs used has not been analyzed yet.

(3) **To test** {whether ~ / this hypothesis/…},
To evaluate X; **To address** the question of X

The second type of PBM is observed in the R-section. As illustrated in (3), it corresponds to a *to*-phrase appearing sentence-initially or finally. It is combined with a description of experimental procedure, as illustrated in (8).

The two types of PBM are both related to a goal description. The first type concerns the whole work and the second type its subset (i.e. an experiment).

3.3 OUTLINE (OTL)

(4) **We report here** the results of experiments…. In brief, we have asked, … To address the first question, we utilized … We found … Together, these results not only confirm that …. but also that… (End of the I-section)

OTL provides a concise characterization of (or an 'excerpts' from) the work as an abstract does.

(5) [Introduction Body Conclusion]$_{\text{full-text article}}$

The rhetorical scheme of the whole article is analyzed as (5). OTL has as its scope "Body", and thus it is expected to appear either in Introduction or Conclusion. This conforms to our investigation.

Tense choices are between simple present and future (in Introduction), and between present perfect and simple past (in Conclusion).

The first element of (4) signals the beginning of an OTL zone. By itself it would fit into AIM (of the paper) employed in (Teufel et al., 2002). It contains certain kind of linguistic signals such as:[3]

(6) Indexicals:
e.g. *in this paper*; *in the present study*; *here*
'Reporting verbs' or verbs for presentation:
e.g. we *show*/ *demonstrate*/ *present*/ *report*

However, OTL consists of a wider range of sentences. As illustrated in (4), OTL also contains those elements which provide information relevant to other zones (e.g. PBM, MTH and RSL). We consider that the whole sequence of sentences in (4) deserve an independent class from both theoretical and practical perspectives. That is, it is embedded in a reporting context, and provides abstract-like information. Thus, we propose OTL.

3.4 TEXTUAL (TXT)

TXT zones were not observed in our sample. This makes sense because the journals investigated provide a rigid section format. However, we retain this class for future application to other journals which may provide a more flexible section format.

3.5 METHOD (MTH)

(7) we **performed** X , using …; we **exploited** the presence of ~; we **utilized** sucrose-gradient fractionation; X **was normalized**

MTH takes the form of an event description in the past tense, using matrix verbs expressing the experimental procedure (e.g. *perform*, *examine*, *use*, *collect*, *purify*). Either a passive or an active form (with *we* as its semantic subject) is used.

(8) [**To test** ~,]$_{\text{PBM}}$ [we **performed** ~]$_{\text{MTH}}$.

We observed that a paragraph in the R-section starts with a combination of PBM and MTH as illustrated in (8). It is much more common for PBM to come first. This can be explained in terms of 'iconicity', the phenomenon that the conceptual and/or the real world ordering of elements is often reflected in linguistic expressions. In (8), the PBM

[2] In what follows, I-, M-, R-, and D- section stand for Introduction, Method and Materials, Results, and Discussion sections, respectively

[3] For a more comprehensive set of expressions, see, for example, (Swales, 1990) and (Teufel et al., 2002).

portion (*to*-phrase) is preposed conforming to the fact that the author first had the experimental goal.

3.6 RESULT (RSL)

(9) the distribution of ~ **was shifted** from …;
no significant change **was seen**;
cells … **demonstrated** an enrichment in ~

RSL usually describes an event in the past tense, as MTH does, using a certain set of verbs expressing; 1) phenomena (e.g. *represent*, *show* and *demonstrate*, having as its subject the material used), 2) observations (e.g. *observe*, *recognize* and *see*, having *we* as its subject, or in the passive form), or 3) biological processes (e.g. *mutate*, *translate*, *express*, often in the passive form).

(10) the distribution of ~ **is shifted** from …
no significant change **is seen**
cells devoid of Scp160p **demonstrates** ~
~ **are presented** in Table 2.

As illustrated above, RSL, unlike MTH, may also be written in the present tense to create a context in which the author observes and presents the results real-time, referring to figures.

In the R-section, RSL zones were observed to follow MTH with no discourse connectives. However, the boundary was rather easy to identify, by virtue of a cause-effect relation identified. Specifically, matrix verbs used in these zones played a critical role; some of them present a rather complementary distribution. This feature is useful for machine learning too.[4]

MTH and RSL may be combined by *resulted in*:

(11) [Parallel … transcription reactions using…] $_{MTH}$
resulted in [… strong smears.] $_{RSL}$

However, *result in* is usually observed in relating biological events, and the above usage relating a method and results is found uncommon. Also, the explicit use of *result(s)* as below is uncommon:

(12) The **results**, ……, were striking. First, …

Given these, keyword searches using *result(s)* do not work for the purpose of identifying experimental results. In contrast, RSL zones can be identified using features such as matrix verbs and location. Thus, annotating RSL zones is important.

(13) Interestingly/ Surprisingly/ Noticeably/…,

In a RSL zone, empathetic expressions as in (13) may be used, often sentence-initially, to call the reader's attention. The adjective version (e.g. *striking* in (12)) is also used.

[4] The occurrences in MTH/ RSL in our sample were: perform 38/2, use 181/12, collect 10/1, purify 23/2, observe 1/43, reduce 1/15, affect 1/15, associate 6/25. However, some verbs had a rather neutral distribution (e.g. detect 11/13, follow 26/8). Such cases require the use of other features too, as we will discuss later on.

3.6.1 INSIGHT (INS)

We have identified three major patterns for INS. The examples below illustrate the first pattern:

(14) [As can be seen in Figure 2C, … was not significantly different compared with that in Figure 2A,] $_{RSL}$ [**indicating that** … had no appreciable effect on ….] $_{INS}$

(15) [Interestingly, central ZYG-9 was significantly reduced in … embryos ……In the converse experiment, … was observed in embryos….] $_{RSL}$
[**These results suggest that** γ-tublin is required to assemble centrosomes ……] $_{INS}$

These are conventionalized forms which the author uses in stating his/her interpretation of the results with respect to a biological process behind the observed results. A generalization is:[5]

(16) X **indicate** Y (a variant: X, **indicating** Y)
X: results/experiments/studies,
Y: biological statement or model,
Verb variations from our sample:
indicate/*suggest*/*demonstrate*/*represent*/*reveal*.

The second pattern is a sentence using the verb *seem*/ *appear* or *consider* such as:

(17) X **seem**/**appear** to V (It **seems**/ **appears** that ~)
X **is considered** to V

The third pattern is the use of *confirm*/ *support*:

(18) This **was confirmed**, as shown in Figure 3.

Here, *this* refers to the author's hypothesis. Although (18) refers to a figure which shows the result, the sentence does not fit into RSL but into INS. We consider that it describes the author's interpretation of the result and that the hypothesis is now licensed as an insight. A generalization is:

(19) X **confirm** that Y; Y **was confirmed**.
X: results/experiments/studies
Y: proposition (hypothesis or prediction).

As we will discuss later, *confirm* also signals CNN, relating two things (X and Y). Therefore, it triggers a nested annotation for INS and CNN.

3.7 IMPLICATION (IMP)

The IMP class is used as a cover category for the author's 'weaker' insights from experimental results and for other kinds of implication of the work (e.g. assessment, applications, future work).

(20) Fusion of …of type III enzymes, …, **would** result in type IIG enzymes…

(21) We **speculate** that as ~ lose …, ~ increases.

'Weaker' insights (vs. 'regular' insights fitting into INS) are signaled by; 1) modal expressions

[5] In our data, *suggest* occurred mainly in INS (63%) and BKG (23%), and *indicate* in INS(55%), RSL(20%) and MTH (10%). This means that these verbs strongly signal INS but other features are also needed for ZI (e.g. location, zone sequence, and the subject of the verb).

(e.g. *could*, *may*, *might*, *be possible*, *one possibility is that*) and 2) verbs related to conjecture (e.g. *speculate*, *hypothesize*), as in the examples above.

(22) These data **are significant** because …
(23) This approach **has the potential to** increase …
(24) ~ **provides** structural **insights into** ~

Assessment is signaled by weak linguistic clues as illustrated in (22) - (24) above.

(25) Potential targets also **remain to be studied;** we **do not yet know**
(26) **Further experiments will** focus on ~; a **future** study/work/challenge…

Taken out of context, IMP mentioning future work look very similar to PBM as in (25), unless it contains key words such as *future* and *further*, as in (26). The critical feature for the distinction between them is the section in which they appear.

3.8 ELSE (ELS)

We found only few cases of ELS in our data. The following is an example (a naming statement).

(27) …, we **refer to** this gene **as** gip-1 and ~ as …

The lack of ELS zone in our data indicates that the domain of experimental biology has a more established methodology and that the focus is on the experiments and the findings obtained. In other domains where the methodology is less standardized (e.g. computer science), we would expect some essential cases fitting into ELS (e.g. the author's proposal and invention) and thus further elaboration of classes would be needed.

3.9 DIFFERENCE (DFF)

(28) [[These effects are significantly **different from** the effects caused by …]$_{DFF}$] $_{RSL}$
(29) [[Our structural results **differ** somewhat **from** the previous proposal (ref.) and …]$_{DFF}$]$_{INS}$

As in (28) and (29), DFF is signalled by a limited set of vocabulary (mainly, *different* and *contrast* and their variants). Also, as illustrated above, DFF often overlaps with other classes (e.g. INS, IMP, RSL), and therefore involves nested annotation.

3.10 CONNECTION (CNN)

(30) This conservation further **supports** their putative regulatory role in exon skipping.
(31) this peroxide treatment experiment **was consistent with** previous data
(32) The results also **confirm** the recent discovery of MntH … (ref).
(33) This conclusion **was supported** not only **by** … but also **by** …

The CNN class covers statements mentioning consistency (i.e. some sort of positive relation) between data/findings. A generalization is:

(34) X is **consistent with** Y ; X **conform to** Y
X **is {similar to/ same as}** Y **;** X **support** Y
X/Y: previous work, the author's observation, model, hypothesis, insight, etc.

(35) X. **Similarly,** Y. (X/Y: a proposition)

The specific relation mentioned shows a variety (e.g. correlation or similarity; support for the author's own or other's data/ idea/ findings). [6]

Interestingly, we observed more CNN zones than DFF zones in our sample (Mizuta et al., 2004), and we consider that this is not accidental; this asymmetry indicates that biologists put more focus on correlation between two elements.[7]

4 Zone identification -2: elaboration

4.1 Nested zones for complex concepts

The following examples illustrate complex zones motivating nested annotation:

(36) [**[Similar** DNA links *were* also *observed* in the complexes with … (ref.), which show structural **similarities with**….**]** $_{CNN}$] $_{RSL}$
(37) **[** [Previous ^{113}Cd NMR studies on … indicated that zinc plays a catalytic role.]$_{BKG}$ [According to the mechanism we propose, Zn^{2+} plays a crucial role only in….]$_{INS}$ **]**$_{DFF}$ **[Another difference** from the previous proposal is…**]**$_{DFF}$

Sentence (36) provides a result and compares it with other results (boldfaced). Thus, the sentence fits into RSL and CNN simultaneously; it is a case of combined zones, conceptually distinct from indeterminacy between two zones. Sentence (37) illustrates an example of nested zones. The first two sentences fit into BKG and INS respectively. Also, they contrast with each other, with respect to the role which zinc is claimed to play, deserving of DFF as a whole (but there is no explicit clue at this point). The key word in the third sentence, *another difference* (boldfaced) licenses the sentence to DFF and also indicates an element referring to a difference already mentioned. Accordingly the first two sentences will be annotated for DFF.

Precisely speaking, combined zones and nested zones are not identical. But we treat combined zones as a special case of nesting, as two zones having the same scope and an arbitrary ordering. Importantly, nested zones (in a wider sense) are conceptually distinct from ambiguity between two zones; the sentences simultaneously fit into

[6] DFF and CNN classes cover a wide range of relations between data and findings.

[7] This insight was checked with a biologist. This asymmetry also suggests the essential difference between the biology and the computer science domains. In the scheme by (Teufel et al., 2002) focusing on computer science articles, CONTRAST seems to be more important than BASIS.

multiple zones. In fact, in our sample, most CNN and DFF zones overlap with another zone such as INS and IMP. Since CNN and DFF zones are important for our purpose, we consider that nested annotation is necessary.

4.2 Controversial cases

(38) **However, it was not evident whether** DPCs composed of … were … or protelytic degradation was involved in the process.

A PBM zone (in I-section) and an IMP zone describing future work (or limitations) often look very similar on the surface, as illustrated in (38), which is the last sentence in the article describing the limitation of the work presented. A critical feature is the location; PBM in this use is located in the I-section, whereas IMP in other sections.

A PBM zone in I-section (e.g. *X remains unclear*) is considered to be a subset of a larger BKG zone when the problem mentioned is a generally accepted fact. However, we chose to avoid nested annotation in this case, because; 1) the situation above is rather common, and yet 2) we identify the significance of PBM zone in its own. In case a single sentence consists of a clause fitting into BKG and another fitting into PBM, then it will result in a complex annotation. That is, we annotate the sentence as both BKG and PBM.

5 Zone identification -3: location

We now analyze the zones appearing in each section and their sequence, to try to describe the locations where a specific zone class may appear.

The section organization of the sample articles is mapped onto the scheme shown in (5) as follows:[8]

(39) $[I_{Intro}\ [M\ R\ D(non\text{-}final)]_{Body}\ D(final)_{Conc}]$

In what follows, I, M, R, and D stand for the corresponding section.

5.1 I-section and M-section

Common to all sample articles, the I-section consists of a large number of BKG zones with a few PBM zones inserted in it, which is then followed by an OTL zone. The OTL zone may or may not constitute a separate paragraph.

The M-section focuses on methodological details, and thus consists of MTH zones possibly with an ignorable number of other zones (e.g. BKG, INS).

5.2 R-section

The R-section consists of 'problem-solving' units following the experimental procedure. The main elements of each unit are PBM, MTH, and RSL zones, which are often then followed by an INS zone. There are also some optional elements. A generalization of the zone patterns is as follows. For practical reasons, we use the regular expression style; superscripts + and * stand for the occurrence of one (+) / zero (*) or more times. Brackets represent OR-relation.

(40) $(X^{*}\ PandM\ MTH^{+}\ (RSL\ INS^{*}\ IMP^{*})^{*})^{+}$

X: an arbitrary zone, and
PandM = [(PBM MTH) (MTH PBM)]

Below are examples of an optional zone (X) placed at the beginning a problem-solving unit:[9]

(41) **[**It is possible that … **]IMP** [To test *this* possibility,]PBM [we examined …] MTH

(42) **[**... has revealed two motifs (Fig. 1). As can be seen in Figure 1A, …….**]RSL** [To ascertain that ….]PBM [we aligned their weight …]MTH

5.3 D-section

The D-section is much more complex and flexible, but some generalization is possible.

First, the essential components of D-section, both quantitatively and qualitatively, are INS and IMP zones. This indicates that the focus of D-section is on obtaining deeper insights. In contrast with the zone sequence in the R-section, INS and IMP often precede, or even lack, RSL and BKG zones related to them. A closer look at examples explains the apparent lack of RSL/BKG:

(43) *The data within this report* demonstrate…

(44) As for the C-rich element, *its comparison with the PTB binding motif* has shown that these are different motifs.

(45) Similarly, *the failure of ... protein (Fig. 7)* suggests that...

The italicized elements in (43) - (45) would fit into RSL or MTH, but are too small constituents to be annotated. As a result, only the whole sentence gets annotated as INS. A similar tendency holds also for BKG (e.g. *since*-clause), but less frequently. We may consider extracting these cases in future work. Usually D-section ends with OTL (summary) or IMP (assessment or future work).

6 Zone identification using multiple features

Table 1 illustrates multiple features contributing to ZI, as we identified them through our manual annotation. We observed that certain pairs of zone classes present similar distribution of key features, with the same primary feature, and that BKG lacks a key feature, indicating its neutral nature. Using multiple features is critical in ZI. We intend to

[8] Or, $[I_{Intro}\ [R\ D(non\text{-}final)\ M\ ]_{Body}\ D(final)_{Conc}]$

[9] We observe that these paragraph-initial zones trigger the PBM zone. For example, *this* in (41) refers to the preceding IMP zone, and the RSL in (42) mentions the results of a preceding experiment.

improve our insight shown here through quantitative analysis (cf. fn. 3 and 4). It then better helps determine the right set of features and their relative priority to be used in machine learning.

Feature\ **Zone**	**B**	**P**	**O**	**M**	**R**	**INS**	**IMP**	**CNN**	**DFF**
lexical/syntactic	-	◯	◯	-	-	◎	◎	◎	◎
matrix verb	-	△	△	◎	◎	-	-	-	-
location	-	◎	◎	◯	◯	△	◯	△	△
zone sequence	-	◯	◯	◯	◯	△	△	△	△
reference to Fig	x	-	-	△	◎	x	x	x	x
citation	◯	-	-	-	x	x	x	△	△

Table 1: Multiple features for ZI

Explanatory notes on the priority of features:
◎: primary feature (with specific clues);
◯: major feature; △: secondary feature
x: negative feature; -: non-/less informative

7 Conclusion

We have provided a qualitative analysis of the process and results of ZI based on our hand-annotated sample, with a view to strengthening the basis for the annotation scheme. We are now starting to use our sample as training data for machine learning, as well as creating more data in a systematic way, toward automatic annotation.

We are also considering to use our ontology management tool (Open Ontology Forge, http://research.nii.ac.jp/~collier/resources/OOF/index.htm) for these purposes; 1) to define zone classes as ontology classes; zone annotation is then expected to be a variant of named entity annotation, which we are familiar with, and 2) to link between expressions referring to results (e.g. *these results*/ *our results*) and their antecedent (i.e. the RSL zone providing a concrete description of the experimental results), using the coreference tool.

Applications include full color representation of annotated texts; a sample is available at: http://research.nii.ac.jp/~collier/projects/ZAISA/index.htm. Also, IR focusing on particular zone classes should improve the quality of retrieval. Specifically, the goal of the experiment (a PBM zone) is expected to be used as an index for the organization and retrieval of experimental results.

Acknowledgements

We gratefully acknowledge the kind support of our colleague Tony Mullen with the quantitative analysis of the data, the generous support of Professor Asao Fujiyama (NII) and the funding from the BioPortal project, and the very helpful comments from the three anonymous reviewers.

References

M. Craven and J. Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In Proceedings of the 7th Intl. Conference on Intelligent Systems for Molecular Biology (ISMB'99).

D.K. Farkas. 1999. The logical and rhetorical construction of procedural discourse. *Technical Communications*, 43(1): 42-53.

K. Humphreys, G. Demetriou and R. Gaizauskas. 2000. Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. In Proceedings of the 5th Pacific Symposium on Biocomputing (PSB2000).

A. Lehman. 1999. Text structuration leading to an automatic summary system. *Information Processing and Management*, 35(2):181-191.

W.C. Mann and S.A. Thompson. 1987. Rhetorical structure theory: toward a functional theory of text organization. *Text*, 8(3):243-281.

D. Marcu and A. Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

Y. Mizuta and N. Collier. 2004. An Annotation Scheme for a Rhetorical Analysis of Biology Articles. In Proceedings of the Fourth Intl. Conference on Language Resources and Evaluation (LREC2004).

C.D. Paice and P.A. Jones. 1993. The identification of important concepts in highly structured technical papers. In Proceedings of the 16th Intl. ACM-SIGIR Conference on Research and Development in Information Retrieval.

J. Swales. 1990. *Genre analysis*. Cambridge UP.

L. Tanabe and W. Wilbur. 2002. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18:1124-1132.

S. Teufel, J. Carletta and M. Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In Proceedings of the 9th EACL Conference.

S. Teufel and M. Moens. 1999. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In "Advances in automatic text summarization", Mani, I. and Maybury, M.T, eds. Cambridge, MA: MIT Press.

S. Teufel and M. Moens. 2002. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28(4):409-445.

T. A. van Dijk. 1980. *Macrostructures*. Hillsdale, NJ: Lawrence Erlbaum.