An Annotation Scheme for a Rhetorical Analysis of Biology Articles

Yoko Mizuta and Nigel Collier

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, JAPAN
{ymizuta, collier} @nii.ac.jp

Abstract

In information extraction from scientific texts, it is crucially important to identify the unique contribution of the research. The task is complicated by the large number of statements made in each article that pertain to results, including reference to previous work and technical details. Simple keyword searches are helpful for a content-based analysis but fail to tell new results from other ones. We aim to approach the problem from a rhetorical perspective and give a 'zone analysis' (ZA) of texts in light of Teufel, Carletta & Moens (1999). We analyze a text into 'zones' with a shallow nesting based on the rhetorical status which each sequence of statements fit into and annotate the text correspondingly. Our current focus is on the molecular biology domain. In this paper, we propose an annotation scheme for ZA based on an empirical analysis of major online journals (EMBO, NAR, PNAS, and JCB), and illustrate how it works. Our scheme provides a way to differentiate the text in terms of the aspects of the author's own work (e.g. experimental procedure, findings, implications) and to identify a set of statements relating data and findings and therefore helps identify the author's new results and findings.

1. Introduction

In information extraction (IE) from scientific texts, it is of critical importance but is not easy to identify the main contribution of the research. To take the example of the molecular biology domain, an immense volume of experimental results have been reported, most of which still remain in the full-text format. It is important to store, utilize, and update major research results in databases. The need is demonstrated by the recent intensive research done on text mining in this domain (Craven et al., 1999; Humphreys et al., 2000; Tanabe et al., 2002). Now we are facing a challenge. The task of identifying new results is complicated by the large number of statements made in each article that pertain to results, including reference to previous work as well as technical details and conjectures. Simple keyword searches are helpful for a content-based analysis but are not powerful enough, because it fails to tell new results from other ones. The same statement may be made *as* a new result, *as* a previously known result, *as* a conjecture, etc., that is, in different rhetorical contexts （e.g. Farkas, 1999）. We need to focus on those statements made *as new results and findings*. We therefore expect that a rhetorical analysis should provide an insight into how our problem can be solved.

There are two lines of approach to rhetorical analysis; Rhetorical Structure Theory (RST: Mann & Thompson, 1988; Marcu et al., 2002) and Zone Analysis (ZA) as we call it (e.g. Teufel, Carletta & Moens, 1999; Hernandez et al., 2003; Farkas, 1999; Lehman, 1999). RST investigates local discourse relations between sentences (e.g. explanation, contrast, elaboration), in a hierarchical fashion, whereas ZA investigates the global rhetorical status of each sentence in terms of argumentation and intellectual attribution. Teufal & Moens (2000) propose analyzing the text into rhetorical 'zones' in a flat structure (e.g. AIM, BACKGROUND, OWN) and provide an annotation scheme. They report the feasibility of the approach based on an application to automatic text summarization of computer science articles and from a machine learning perspective.

We find the latter line of approach, ZA, suits our purpose, since we are concerned with filtering the text according to the relevance to the author's results and findings. We thus attempt to apply ZA to IE. Specifically, we aim to annotate the text based on ZA. Our current focus is on the molecular biology domain. While we aim at a generalization to a larger scientific domain, application to the biology domain deserves our attention in its own right for the reasons mentioned above.

As the first step, we investigated a total of 20 biology articles taken from four major online journals (i.e. EMBO: European Molecular Biology Organization, NAR: Nucleic Acid Research, PNAS: Proceeding of National Academy of Science, and JCB: Journal of Cell Biology), using an existing annotation scheme mentioned above (Teufel, Carletta and Moens 1999: henceforth, 'TCM'). Despite the relatively small number of articles analyzed, our analysis exemplified some critical issues to be considered in the annotation scheme.

In this paper, we first summarize our empirical analysis of online journal articles, and then propose an annotation scheme for biology articles, illustrating how it works. The proposed scheme provides a way to differentiate the text in terms of the aspects of one's own work (e.g. experimental procedure, findings, implications) and to identify a certain set of statements relating data and findings the author's own or to others. We conclude that this is a step forward in identifying the author's new experimental results and findings.

2. Requirements for the annotation scheme

The journals investigated have a common section format, consisting of "Introduction", "Materials and Methods", "Results", and "Discussions" sections (henceforth, the I-, M-, R-, and D- sections) together with an Abstract.[1] The articles investigated fit into an experimental framework and each section has its own pattern of argumentation shared by most articles. The whole article is committed to a main problem-solving such as 'the identification of the effect of Mnt deletion in governing key Myc functions' and 'the identification of the receptor for MAG that elicit morphological changes in neurons', which is done as a

[1] While some authors combine the Results and Discussion sections, here we basically assume an uncombined version.

combination of smaller units. A generalization is this: the I-section introduces the problem and outlines the content of the article, the M-section states the methodological details, the R-section states smaller problem-solving units, and the D-section synthesizes the results and findings and provids prospective remarks. To be emphasized, important information is provided across sections, from a broader perspective in the I- and the D- sections, and from a more specific perspective in the R-section.

Through our empirical analysis of the four journals, we identified the following issues that need to be considered in the annotation scheme. The first two concern the design of zone classes and the third one concerns the annotation principle.

Fine-grained classification of one's own work

The following passage taken from the R-section (NAR, 2003, 31(7), p.1871) illustrates a problem-solving unit (numbering is ours and dots indicate removed words):

> [1] Microarrays have been used to map replication in yeast (ref.). [2] We performed a similar experiment in Salmonella [3] to demonstrate that …..... , [4] First, …... [5] The resulting plot represents the relative increase in gene copy number ……(Fig.). [6] A similar experiment was performed using …… (Fig.). [7] The position of genes …… are scrambled …... [8] We speculate that …… [9]…, this experiment shows that …… [10] The results also confirm the recent discovery of MntH …… (ref.)

The statements above provide various rhetorical information; background information ([1]), the goal of the experiment ([3]), experimental procedures ([2], [4], [6]), results ([5], [7]), and the author's interpretations of the results ([8] -- [10]). Since the R-section consists of such units, the author's new results and findings (e.g. [8] -- [10] above) spread over the section.

The TCM scheme provides a single class "OWN", which covers virtually all aspects of the author's own work.[2] Thus, sentences [2] through [10] above would all fit into OWN. For our purpose we have found it necessary to develop a more fine-grained classification of the author's own work sensitive to the aspects of the work as illustrated above.

Relation between data and findings

The following are the examples of commonly-found statements relating the author's own data and findings to their own or others' (NAR, 2003, 31(7); italics are ours):

(1) 'this peroxide treatment experiment *was consistent with* previous data'

(2) '*The results also confirm the recent discovery* of MntH as an important component of H_2O_2 resistance and virulence in Salmonella (ref).

Such statements indicate the (in)consistency between data and findings and therefore are worth annotating.

The TCM scheme offers two zone classes which are relevant here, CONTRAST and BASIS. They are, however, used for rather restricted purposes: they identify the author's (positive or negative) attitudes toward other work as well as the status of the author's work in research paradigms (Teufel & Moens, 2002). We need to cover a wider range of relations concerning data and findings

Nested annotation

Another issue concerns the annotation principle. The example below (EMBO, 2003, 22(20)) illustrates a case motivating nested annotation (italics are ours):

(3) the average length of the new flagellum *was shorter than* the one measured from replicating slender cells at the same stage (*21 µm instead of 25 µm*) (ref.).

Here, the whole sentence simultaneously; 1) states the author's result, and 2) compares it with previously provided data. Thus, the sentence fits into complex rhetorical classes. This is conceptually distinct from ambiguity between two classes and from a combination of clauses fitting into different rhetorical classes. It thus motivates combined (more generally, nested) annotation.

For Teufel and Moens' (2002) scheme, zone classes should be non-overlapping and annotation should be unique. For example, they suggest having a smaller annotation unit in cases where the sentence consists of clauses fitting into different classes. However, the example above illustrates a new case. Our intuition tells that it is one thing for classes to be conceptually non-overlapping and that it is another for a linguistic unit to fit into a single class. That is, a linguistic unit may well represent complex concepts. Therefore, we consider that nested annotation is a necessity, even though it complicates annotation

3. Annotation scheme for biology articles

Based on the issues discussed in the last section, we provide below an annotation scheme for biology articles.

The set of zones

The set of zones is as follows:

- BACKGROUND (BKG): an assumption referring to previous work or a generally accepted fact.
- PROBLEM SETTING (PBM): a problem to be solved and/or the goal of the present work/paper.
- OUTLINE (OTL): a characterization or a summary of the content of the paper.
- TEXTUAL (TXT): the organization of the paper.
- OWN: the author's own work. Sub classes:
  - ◊ METHOD (MTH): methodology and materials;
  - ◊ RESULT (RES): the results of the experiment performed;
  - ◊ INSIGHT (INS): the insights/findings obtained (e.g. the author's interpretation of the result);
  - ◊ IMPLICATION (IMP): the implications of the experimental results, including conjectures, assessment, applications, and future work;
  - ◊ ELSE (ELS): anything else about the author's work.
- DIFFERENCE (DFF): a contrast or inconsistency between data and/or findings.
- CONNECTION (CNN): a relation or consistency between data and/or findings.

[2] Their earlier classification was more fine-grained in terms of problem-solving (e.g. SOLUTION/METHOD, RESULT, CONCLUSION) but led to rather indeterminate annotation (Teufel and Moens, 1999). Their later version with the single OWN class (Teufel and Moens, 2000; 2002) focuses on intellectual attributions among researchers.

The fine-grained classification of OWN with the five subclasses makes it possible to focus on the author's findings annotated as INS, and on other aspects of the work. The DFF and the CNN classes cover a certain set of statements, which mention what supports (or refutes) a certain finding or data, and which indicate the (in)consistency between data provided in researches. Examples (1) and (2) both fit into CNN.

The OTL class is another unique element in our scheme. It may provide the aim, the goal, the main results, and/or methodological descriptions, and therefore may well overlap with the information fitting into other classes. We consider though that it is deserving of an independent class, as it provides a concise characterization of the work or an 'excerpts' from the work as an abstract does.[3]

Annotation principle

An annotated zone may be as small as a phrase and as large as a paragraph, depending on linguistic and other features. To take examples of a smaller zone:

(4) **To test whether ….., ** we performed ….

(5) X (: a list of experimental results)**, indicating that Y** (: a statement of a finding)

The boldfaced phrase in (4) signals the goal of the experiment and is annotated as PBM, whereas the clause in (5) signalled by 'indicating that' is annotated as INS. In contrast, a sequence of methodological descriptions as observed in the M-section constitutes a larger MTH zone.

We also employ nested annotation on an empirical and theoretical basis mentioned above. This helps reduce the unavoidable inconsistency of a flat annotation scheme noted by Teufel & Moens (1999): we attribute it to the non-clear-cut nature of categorization in general rather than to misclassification. To control the human error caused by a complex annotation scheme, we only allow for one-level nesting. Example (3) fits into INS and CNN.

4. Application of the scheme

Based on the issues discussed in the last section, we provide below an annotation scheme for biology articles.

Sample annotation

The passage mentioned in Section 2 is annotated as shown in Figure 1. The last part illustrates a nested annotation, with a CNN zone embedded in an INS zone.[4]

Zone identification

Based on a sample of hand-annotated data, although rather small in amount at this point, we discuss our preliminary investigation on zone identification.

In the R- section, it is common for a MTH and a RSL zone to appear in pair, as shown in Figure 1. They present a complementary distribution in matrix verbs, and therefore identifiable; MTH takes verbs related to an experimental procedure (e.g. *perform*, *examine*, *use*), whereas RSL takes verbs related to phenomena and observation (e.g. *represent*, *show*, *observe*) and biology-specific verbs in the passive form (e.g. *be + scrambled*, *transformed*).

<BKG> Microarrays have been used to map replication in yeast (ref.).

<MTH> We performed a similar experiment in Salmonella

<PBM> to demonstrate that …....

<MTH> First, …...

<RES> The resulting plot represents the relative increase in gene copy number ……(Fig.#).

<MTH> A similar experiment was performed using …… (Fig.#).

<RES> The position of genes …… are scrambled …...

<IMP> We speculate that ……

<INS> …, this experiment shows that ……

<CNN> The results also confirm the recent discovery of MntH …… (ref.)

Figure 1: Sample annotation of a passage in the R-section (dots indicate removed content words.)

An INS zone, either in the R-section or in the D-section, is usually signalled by 'indicate that' or the like (e.g. *suggest*, *demonstrate*, *reveal*, *show*), having the results or experiments as the subject:

(6) These(/Our) results **indicate that** Y (: a finding).

This is a conventionalized form which the author uses in stating his/her interpretation of the results. Its variant illustrated in example (5) is also commonly used. An alternative form signalling an INS zone (following a RSL zone) is 'Therefore, …… seem/appear to …'.

An IMP zone is currently used as a cover category for the author's 'weaker insights' from experimental results and for any kind of implication of the research, including applications and future work. Thus, it still leaves room for conceptual sophistication. 'Weaker insights' as opposed to 'regular' insights fitting into INS are signalled by modal expressions (e.g. *could*, *may*, *might*, *be possible*, *one possibility is that*) and verbs related to conjecture (e.g. *speculate*, *hypothesize*).

Zone distribution

Table 1 shows the distribution of zones, investigated using four articles taken from the NAR journal..

The section-by-section zone distribution is given vertically in the 'A' columns. For example, 74.4% of the I-section fit into a BKG zone. The zone distribution across sections is given horizontally in the 'B' columns. For example, 68.7% of BKG zones appeared in the I-section. Calculation was done by the number of words.[5] The 'A' column totals exceeding 100 is due to nested annotations. Common cases of nested annotation were; CNN and DFF zones embedded in an INS or an IMP zone (in the R- and D-sections), and PBM zones overlapping with a BKG zone (in the I-section)

[3] Whereas the abstract is provided outside the full text, OTL is provided outside the 'body' of the full text (in the I-section).

[4] We use a graphical representation of annotation, where zones are indicated in different colors (and fonts, for DFF and CNN).

[5] Subsection titles are exempt from annotations but are included in the total number of words in the section. This explains the 'A' column total for the M-section being below 100.

	I-section		M-section		R-section		D-section		Total
Zone	A	B	A	B	A	B	A	B	**of B**
BKG	74.4	68.7	1.1	1.7	4.4	10.4	12.2	19.2	100.0
PBM	5.1	19.7	0.0	0.0	4.8	47.3	5.0	33.0	100.0
OTL	25.1	81.8	0.0	0.0	0.0	0.0	3.3	18.2	100.0
MTH	0.0	0.0	96.7	57.5	39.2	36.9	8.7	5.5	100.0
RSL	0.0	0.0	0.0	0.0	38.9	77.7	16.5	22.3	100.0
INS	0.0	0.0	0.8	4.3	5.7	45.7	9.2	50.1	100.0
IMP	0.0	0.0	1.0	1.4	9.1	20.3	52.0	78.2	100.0
CNN	1.4	5.1	0.0	0.0	3.2	29.5	10.5	65.5	100.0
DFF	3.0	29.4	0.0	0.0	2.8	70.6	0.0	0.0	100.0
Total	109.0	n.a.	99.7	n.a.	108.2	n.a.	117.4	n.a.	n.a.

Table 1: The distribution of zones (% by the # of words) within each section (A) and across sections (B)

As shown in the 'B' columns, most part of the INS, IMP, CNN, and DFF zones appeared in the R- and the D-sections. Table 1 mostly conforms to our observations of all articles studied. An idiosyncrasy of this sample is that DFF zones are few and are missing in the D-section.

The 20 articles studied indicate the following. The R-section presents a regular pattern of zone sequence (i.e. PBM-MTH-RES-INS/IMP for each problem-solving unit), as shown in Figure 1. The D-section presents quite flexible patterns across articles. Characteristically though, the D-section contains larger IMP zones, each of which provides complex content including conjectures and arguments toward deeper interpretations of the results.

5. Theoretical and practical implications

We have proposed an annotation scheme which helps identify the author's results and findings and a certain type of relations between data and findings. To identify the relative significance of information provided within a zone or across zones, further analysis in line with RST would be helpful. However, a rather small set of relations would suffice including 'cause/explanation' (signalled by *therefore* etc.), 'opposition' (signalled by *however* etc.), and 'list' (AND and OR relations)[6].

We expect that our annotation scheme could be applied also to other domains in an experimental framework (e.g. experimental physics and psychology) with minor modifications, if any. Articles in a theoretical framework would require some substantial modifications of the scheme; for example, the author's proposal and its effect should make important zone classes

6. Concluding remarks

The proposed scheme provides a way to differentiate the text in terms of the aspects of the author's own work (e.g. experimental procedure, findings, implications) and to identify a certain set of statements relating data and findings. We therefore conclude that it is a step forward in identifying the author's unique experimental results and findings. As a future direction, we aim to collect a larger number of hand-annotated samples and use them as training data for machine leaning. We then aim at automatic annotation in the proposed scheme for further steps toward our goal.

Acknowledgements

We are grateful for the helpful comments from two anonymous reviewers.

References

Craven, M. & Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In Proceedings of ISMB'99 (pp. 77--86).

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

Hernandez, N. & Grau, B. (2003). What is this text about? Combining topic and meta-descriptors for text structure presentation. SIGDOC'03.

Humphreys, K., Demetriou, G. & Gaizauskas, R. (2000). Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. BSB 2000.

Hyland, K. (1998). Persuasion and context: The pragmatics of academic metadiscourse. Journal of Pragmatics, 30, 437--455.

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

Mann, W.C. & Thompson, S.A. (1988). Rhetorical structure theory: toward a functional theory of text organization. Text, 8(3), 243--281.

Marcu, D. & Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. ACL2002.

Paice, C.D. & Jones, P.A. (1993) The identification of important concepts in highly structured technical papers. ACM-SIGIR.

Shah, P. K., Perez-Iratxeta, C. Bork, P. & Andrande, M. A. (2003). Information extraction from full text scientific articles. In BMC Bioinformatics, 4--20.

Swales, J. (1990) Genre analysis. Cambridge Univ. Press.

Tanabe, L. & Wilbur, W. (2002). Tagging gene and protein names in biomedical text. In Bioinformatics, 18, 1124--1132.

Teufel, S., Carletta, J. & Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. EACL '99.

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

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

Teufel, S. & Moens, M. (2000). What's yours and what's mine: Determining Intellectual Attribution in Scientific Text. SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

Tischer, S., Meyer, M., Wodak R., and Vetter E. (2000). *Methods of text and discourse analysis*. London: SAGE Publications Ltd.

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

[6] Marcu & Echihabi (2002) propose a simplified version of RST classes (i.e. 'contrast', 'cause-explanation-evidence', 'condition', and 'elaboration') from a machine learning perspective.