A baseline feature set for learning rhetorical zones using full articles in the biomedical domain

Tony Mullen
National Institute of Informatics
2-1-2, Chiyoda-ku
Tokyo 101-8430, Japan
mullen@nii.ac.jp

Yoko Mizuta
National Institute of Informatics
2-1-2, Chiyoda-ku
Tokyo 101-8430, Japan
ymizuta@nii.ac.jp

Nigel Collier
National Institute of Informatics
2-1-2, Chiyoda-ku
Tokyo 101-8430, Japan
collier@nii.ac.jp

ABSTRACT

At a time when experimental throughput in the field of molecular biology is increasing, it is necessary for biologists and people working in related fields to have access to sophisticated tools to enable them to efficiently process large amounts of information in order to stay abreast of current research.

Rhetorical zone analysis is an application of natural language processing in which areas of text in scientific papers are classified in terms of argumentation and intellectual contribution in order to pinpoint and distinguish certain types of information. Such analysis can be employed to assist in information extraction, helping to assess and integrate data generated by experiments into the scientific community's store of knowledge.

We present results for several experiments in automatic zone identification on the ZAISA-1 dataset, a new dataset composed of full biomedical research papers hand-annotated for rhetorical zones. We concentrate on general purpose and linguistically motivated features, and report results for a variety of sets of features. It is our intention to provide a baseline feature set for modeling, which can be extended in future work using combinations of heuristics and more sophisticated and task-specific modeling techniques.

1. INTRODUCTION

As increasing amounts of new data become available as the result of high throughput experiments in molecular biology, it becomes ever more important to have sophisticated methods for automatically extracting, assessing, and integrating new information from published papers into forms easily accessible to the scientific community. Information extraction (IE) is the task of mapping information from unstructured natural language texts such as journal articles to partially structured representations of meaning suitable for databases. This is now regarded as a fundamental technique for utilizing information contained in archived journal articles and abstract collections such as MEDLINE, but there is much room for improvement. Major progress in natural language processing (NLP) approaches related to this task has been made by researchers in the extraction of biological named entities, i.e. the identification and classification of technical terms such as proteins, genes, drugs, or cell types, and biological interactions [5; 8; 20]), but further progress aimed at pin-pointing and organizing factual information remains a challenge [7].

A step towards more sophisticated IE can be taken by analysis and exploitation of *rhetorical zones* in a scientific text. The task of zone analysis (ZA) involves classification of areas of text, or *zones* in terms of argumentation and intellectual attribution. Teufel [22] proposed an analysis of text into rhetorical zones in a flat structure and provides an annotation scheme which is then applied to text summarization in the domain of computational linguistics [23; 24]. Mizuta and Collier [13; 15] propose the application of ZA to biology articles from an IE perspective. Other areas in which zone analysis may be of use are information retrieval (IR) and citation analysis[13; 17].

In this paper, we present the results of machine learning experiments on zone analysis in the biomedical domain using the ZAISA-1 dataset annotated according to the version of the guidelines published in [14].

1.1 The need for zone analysis

Pin-pointing and organizing factual information is an important part of any IE task, but this consists of more than merely identifying factual statements within the texts. Contextual information is often necessary to understand how to interpret the statements contained in sentences such that the pertinent factual content can be extracted. For example, a proposition in itself may refer to a current result or an old result, it may refer to a proposed theory or a disproven theory, it may refer to a description of the outcome of an experiment, or a description of the expected outcome. The distinctions between these rhetorical contexts are crucial for correctly interpreting the meaning of the research. Furthermore, the necessary context information to interpret a fact may be subtle, and is likely to occur outside of the sentence in which the fact is found.

In cases such as the following sentences, information which could be easily mistaken for results without knowledge of rhetorical zones is clearly intended as background (**BKG**) or problem (**PBM**) information, according to the zone annotations.

<**PBM**>*Surprisingly, however, each of these mice is viable, fertile and fail to spontaneously give rise to tumors, an outcome predicted for a Myc antagonist.*</**PBM**>

<**BKG**>*However, incubation of these cells at intermediate temperatures between restrictive and permissive temperatures leads to flagella of intermediate lengths (Marshall and Rosenbaum, 2001).*</**BKG**>

<**BKG**>*However, the CAS/Cse1p binding site on human importin- maps to the convex face of Arm repeats 9 and 10 (Herold et al., 1998).* </**BKG**>

Sentences and phrases which are ambiguous without proper recognition of their rhetorical context are frequently used. As pointed out in [19], sentences themselves contain propositions, but texts are more than simple bags of sentences. Rather, contextual information is encoded in the *coherence relations* between sentences, and the facts contained in the sentences should be interpreted within the discourse. Human annotation of rhetorical zones provides metatextual support for recognizing specific types of information and exploiting discourse context.

At present, a considerable amount of research in IE for biomedical texts has focused on abstracts [11; 10; 16; 8]. The target of biological information extraction should arguably be full texts, however, given their much richer sources of information and the increasing ease of access in the form of online journals. Our approach to zone analysis aims to capitalize on the richness of information available in full texts.

2. METHODS

2.1 Zone Analysis

The goal of zone analysis is to assign areas of text to (possibly overlapping) rhetorical zone classes [23]. These zone classes vary in the nature of information they are intended to convey. The specific classes considered in this work follow those in [13; 15]. Zones are furthermore grouped into three separate groups, more broadly representing the kinds of information conveyed [14]. The set of zones is as follows:

- **Group One**
  - 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
  - OWN: the author's own work:
    - MTH (Method): methods used
    - RSL (Result): experimental results
    - INS (Insight): direct, objective interpretations of the data in experimental results
    - IMP (Implication): broader implications of experimental results, e.g. conjectures, assessment, applications, future work
    - ELS (Else): anything else within OWN.
- **Group Two**
  - CNN (Connection): correlation or consistency between data and/or findings
  - DFF (Difference): a contrast or inconsistency between data and/or findings
- **Group Three**
  - OTL (Outline): the summary of the paper

The first group includes zones whose information makes up the main content of the paper. The second group contains zones whose information compares or contrasts data or findings from the paper with other data or findings. The third group is composed a single class, which conveys meta-textual information about the paper and its organization.

In the present work we concentrate on Group 1. The **OWN** class is exceptional in that zones are not classified directly as **OWN** but must be classified further as one of the subclasses of **OWN**. So for the task of classification we can ignore **OWN** itself. Furthermore we disregard **ELS** in the present experiments due to its rarity (only 6 sentences contain **ELS** annotations) in the current dataset.

An important question from both a human annotation and a machine learning standpoint is what the basic units of classification should be. According to the annotation guidelines in [14], the unit of annotation may be either a sentence or a phrasal constituent, with sentence-scope annotations being more common. In the current work, however, it is necessary to decide on a single discrete unit of classification, so we define our task as classification of sentences, and we do not analyze phrasal constituents as such. We consider a sentence to belong to a class if the sentence or any constituent within the sentence is annotated as belonging to the class. Thus, if a sentence contains a phrase which has been annotated as a **RSL** phrase, we consider the sentence to be a positive instance of the **RSL** class. In this sense, the classification approach taken here is coarser than the analysis specified in the annotation guidelines. In the current version of the guidelines, overlapping zones are allowed, so a sentence may belong to several sentence-scope zone classes in addition to containing constituent-scope zones.

In most cases, accurately classifying sentences is enough to accurately annotate the zones of a text. However, in cases where sub-sentential phrases are highlighted by a human annotator, a deeper level of analysis is required. In the present experiments, we take the approach of classifying sentences. In future work, this step may be used as a first pass, in cases where sub-sentential analysis is warranted. In this case a second pass aimed at classifying the sub-sentential constituents within multiply classified sentences, can be employed.

2.2 Classification methods

We classify sentences using binary classifiers, with a separate classifier representing each class. The classifiers for the various classes are independent of each other. We experimented with two well-known methods of classification, Naive Bayes and Support Vector Machines. Both classifiers are trained using labeled training examples in the ZAISA-1 dataset.

2.2.1 Naive Bayes

Naive Bayes is a general purpose method of machine learning widely used for tasks such as text classification. In this method, a simple statistical classifier is created based upon Bayes' Law, notable for fast modeling and low computational overhead. Its primary drawback is its assumption of statistical independence of the features used for modeling, making it less able to take advantage of the information available in complex feature sets containing inter-dependent features.

Sentence	Relation	Dependency triple
John ate the apple	subj	john_eat_subj
John ate the apple	obj	appole_eat_obj
John wanted to eat	v-ch	eat_want_v-ch
John sat in the park	pcomp	park_in_pcomp

Figure 1: Simple examples of grammatical relations between words represented as dependency triples.

2.2.2 Support Vector Machines

SVMs are a machine learning classification technique which use a function called a *kernel* to map a space of data points in which the data is usually not expected to be linearly separable onto a new space in which it is, with allowances for erroneous classification [25]. SVMs are known for high performance, particularly in tasks with large numbers of features. SVMs make no assumption of statistical independence, which makes them more suitable than Naive Bayes in cases where features may contain redundant, overlapping, or otherwise interdependent information, as is often the case in complex feature sets. For a tutorial on SVMs and details of their formulation we refer the reader to [4] and [6]. A detailed treatment of these models' application to text classification may be found in [9].

We use Kudo's TinySVM implementation for our experiments.[1] A polynomial kernel with the degree parameter set to 2 was used.

2.3 Features for zone analysis

A number of types of information are available to be incorporated into the model. [22] gives an extensive list of possibilities for features directly applicable to the task of zone analysis. In the present work, we focus on establishing a baseline of modeling the data available to us, using a collection of general purpose, intuitive features. Since we have formulated the task as a classification task with sentences as its units of classification, we have experimented with traditional text-classification features, including unigrams and bigrams of word tokens and lemmatized words, i.e. the morphologically uninflected form of the word. Rhetorical zones often strongly correspond with certain words and phrases, for example the phrase *we found* would probably suggest results or insights. Although we do not use set phrases per se as features, it is expected that the word and lemma n-grams used should automatically incorporate such phrases into the model. We employ the Conexor FDG parser [21] for lemmatization and grammatical analysis and extract information about the syntactic dependencies of each word, derived from the grammar. This information is encoded in the form of *dependency triples* which include the lemma form of the word in question, the lemma form of the word it is dependent on, and the syntactic relationship between the two words. Examples of dependency relations include verb-subject (`subj`), verb-object (`obj`), main verb-auxiliary verb (`v-ch`), preposition-complement (`pcomp`), and other linguistic relations. Figure 1 shows how these relations are represented as dependency triples.

A standard approach when using word-based models in text classification and IR is to exclude *stopwords*, words judged in advance to be unlikely to yield useful information [18; 9]. These are usually common function words such as the determiners *a* or *the*, pronouns, and certain prepositions. We evaluate the usefulness of excluding a small list of high-frequency function words.

[1] `www.tahoo.org/~taku/software/TinySVM`

3. EXPERIMENTS

3.1 The ZAISA-1 Dataset

The ZAISA-1 Dataset is composed of 20 full journal articles in the area of molecular biology; 5 articles from the European Molecular Biology Organization (EMBO) Journal, 5 articles from the Proceedings of the National Academy of Science (PNAS), 6 articles from the Nucleic Acid Research (NAR) Journal, and 4 articles from the Journal of Cell Biology (JCB). The articles have been hand annotated for rhetorical zone information by a linguist according to the guidelines published in [15].

The full dataset consists of 3637 sentences. Counts of positive examples by class may be viewed alongside the results for machine learning experiments for the corresponding classes in Figure 3. A more detailed breakdown of the correspondence between classes and zones may be seen in Figure 2. In some experiments, only results sections are used. This subset of the data consists of 1727 sentences, 874 of which are positive examples of results.

3.2 Experiments

Two main sets of experiments were carried out, one on full articles, using all classes, and another on results sections only, using only the class RSL. Extracting information about new results is of primary importance in scientific IE and exploiting section knowledge allows us to focus attention to where those results are most likely to occur.

The results in Figure 3 come from 10-fold cross validation, using 2 articles as a test data for each fold and training on the other 18. Although the unit of classification is sentences, the unseen data in actual applications would be full papers, and contextual information such as paper-specific unigrams and information from preceding and subsequent sentences plays an important role in the classification, so we are careful not to include sentences in the training data which come from the same paper as the sentence being classified.

Feature sets for the reported experiments are broadly motivated by the investigation in [13] and are presented here as collections of features representing lexical/syntactic information, information about the main verb of the sentence, information about the location of the sentence in the text, and sequence information. Note also that in Figure 3 features are added cumulatively from left to right, and not taken away. The feature sets in the rightmost columns include the features from the columns to the left.

- **Lexical/syntactic** Composed of features representing word unigrams; lemma unigrams and bigrams; and dependency triples derived from analysis by the parser. Short (two or three word) phrases will be automatically represented as n-grams. Also, all citations are converted to a single token, `CITE`, which is represented here.

- **Main verb** Features representing the main verb in the sentence (as identified by the parser) by lemma, as well as morphological information about the verb including

		Percentage of words in each section by class					
Section	**Total words**	**BKG**	**PBM**	**MTH**	**RSL**	**INS**	**IMP**
ABSTRACT	3,274 wds	17%	13%	7%	44%	17%	1%
INTRODUCTION	13,028 wds	80%	9%	2%	3%	4%	
MATERIALS AND METHODS	17,986 wds	1%		95%		1%	2%
EXPERIMENTAL PROCEDURES	1,215 wds			96%		3%	
RESULTS	41,277 wds	9%	7%	19%	52%	10%	5%
RESULTS AND DISCUSSION	8,489 wds	25%	5%	9%	34%	12%	18%
DISCUSSION	19,655 wds	23%	2%	2%	22%	20%	31%

Figure 2: This table indicates the correspondence between zones and sections. Text from the entire dataset is grouped by section. The names of sections are somewhat variable from paper to paper. For example, some authors prefer to combine Results and Discussion sections into a single section. The left column shows a list of the section headings as they occur in the ZAISA-1 data set. The second column shows the total number of words in each section. The other columns show the percentage of words labeled for each class. Because of overlapping classes, these percents do not necessarily sum to 100.

tense and voice. The main verb is the central predication of a sentence, and contains a considerable amount of information about the meaning of the sentence, and is therefore likely to be useful in many sentence classification tasks.

- **Location** Composed of features representing the name of the section in which the sentence occurs, and the absolute location feature described in [22] in which papers are divided into ten location indices. Sections and zones have a high degree of correspondence as can be seen in Figure 2.
- **Zone sequence** This adds all of the above feature information for previous and subsequent sentences, within a window of +1-1 sentences around the focus sentence being classified, yielding a weakly sequential model.

In cases where a particular feature type has a large number of values, the most frequent 200 features only are considered for each such feature type. For example, a model using word unigrams would consist of 200 features, and a model using unigram and bigram features would consist of 400 features, 200 for each feature type. This is a common method of keeping the models within a reasonable size, and can also be helpful for modeling by limiting noise due to low-frequency features.

In the experiments on full texts, features are extracted from the sentence being classified only, except in the "Zone sequence" experiments, where a +1-1 window is employed; i.e., features are collected from the sentence being classified, as well as from the preceding and following sentences, yielding a weakly sequential model. Likewise, in the experiments on results sections the best models were produced using a +1-1 window.

4. RESULTS

The results of the experiments on whole texts using all classes may be seen in Figure 3. Results of the experiments identifying only RSL classes in known results sections may be seen in Figure 5. Results of experiments specifically comparing the benefits of location information derived from section headings to information derived from absolute location indices may be seen in Figure 4.

On full articles, the highest overall f-score, 70, is obtained by the SVM model using the full feature set. Of the specific classes, recognition of MTH yields the highest f-score by a considerable margin at 87. In almost all cases, adding features yields some improvement in the SVM models, whereas increasing the complexity of the feature set does not benefit the Naive Bayes models as much. Location emerges as a key figure in many cases; the addition of location features yields the sharpest increase in overall f-scores for both models. Contextual information derived from looking at features in a +1-1 window appears to be better exploited by the SVM model than by Naive Bayes.

A comparison between two types of location information suggests that section information is considerably more helpful than absolute location information but that the two types together yield a benefit in overall f-score which is somewhat greater than that gained by either of the two separately. The absolute location index feature yields no improvement in f-score over the model with all other non-location features, but the precision and recall values are somewhat more balanced. Section heading information yields improvement in precision and recall. Use of both features yields improvement in recall. It would appear that the benefits of each information source are somewhat independent of each other, yielding a cumulative improvement when both feature types are included, although the contribution of the absolute location information is too small to be confident of.

In the experiments on results sections only, where the goal was to identify those sentences belonging to RSL classes only, the highest f-score obtained is 80, incorporating a variety of lexical, syntactic, and location features, along with a +1-1 window.

5. DISCUSSION

In general, it appears that some improvement may be gained by almost all the features employed. The SVM models in particular benefit from the larger feature sets. The relative success of the Results sections-only experiments and the improvement gained by the location oriented features in the full text experiments confirm the intuition that actual location within the text is of key importance to the task of zone analysis.

Results on n-gram style sequential models in [22] would appear to give reason to doubt that sequential information can be usefully employed in this task, but as we see from the improvement gained by use of information from the preceding and following sentences, local context does appear to

Class	Sentence count (% of data)	Lexical/syntactic		Matrix Verb		Location		Zone sequence	
		NB	SVM	NB	SVM	NB	SVM	NB	SVM
INS	404 (11%)	41	41	38	43	38	44	35	44
BKG	730 (20%)	41	55	49	55	62	65	61	71
IMP	344 (9%)	41	36	41	34	43	48	43	48
MTH	1063 (29%)	74	76	75	75	81	84	81	87
RSL	1108 (30%)	59	59	59	63	61	68	61	70
PBM	283 (8%)	53	49	56	52	58	57	56	53
Overall		57	59	57	61	63	67	63	70
Overall Prec/Rec		70/48	65/55	68/50	67/56	76/55	68/66	73/55	79/63

Figure 3: F-score and overall precision/recall results for ten-fold cross-validation experiments using full articles from the ZAISA-1 dataset, and all Group 1 classes except ELS. Results are shown for Naive Bayes and SVM binary classification models for each zone class.

Features	Precision	Recall	F-Score
Full feature set without location	77	56	65
Full feature set with location indices	76	60	67
Full feature set with section information	81	60	69
Full feature set with all location information	79	63	70

Figure 4: A closer look at the contribution made by specific location-related features on classification of all zones within full texts. Overall precision, recall, and f-score results are shown.

Features	Prec	Recall	F-Score
Word unigrams only	79	75	77
Word U & B	79	74	76
Lemma unigrams only	79	73	76
Lemma U & B	79	74	77
Dependency triples only	68	47	56
Dependencies, main verb lemmas and morphology	70	60	64
Word U, lemma U & B, deps, MV lems, MV morph	82	76	79
Word U, lemma U & B, deps, MV lems, MV morph, Abs. Loc. index	81	78	**80**

Figure 5: Results for experiments on results sections only, identifying only the single class RSL. Features include word and lemma unigrams and bigrams, dependency triples, main verb lemmas, main verb morphological information, and absolute location indices. Results shown are from SVMs using a +1-1 context window with various combinations of these features.

	Stopwords excluded	All words included
All zones/whole papers	52	55
RSL zones/Results sections only	70	73

Figure 6: F-score results comparing the use of stopwords with unrestricted word token unigram models. In the unrestricted case the top 200 words by frequency are included. In the stopwords case a list of 20 frequently occurring determiners, conjunctions, prepositions and pronouns is excluded, and the top 200 remaining words are included.

contribute to recognition. This also conforms to intuitions about the task. Sentence to sentence classification is not strongly sequential in the same way that word-to-word or letter-to-letter prediction tasks would be, which might explain why a strongly sequential model would not optimally exploit the information available from local context. In the present experiments, the SVM model does a particularly good job of taking feature information from the local window into consideration.

In some cases where full sentences were mis-classified, the correct zone classification could have been identified with slightly deeper linguistic analysis, as in the following example:

In contrast to loss of Mad proteins , Mnt deficiency was found to cause a phenotype remarkably similar to that caused by Myc overexpression and to predispose cells to tumorigenesis in vivo.

which was incorrectly classified by the learner as a result, when in fact the annotator identified the sentence as being an insight. It can be surmised that the learner made the mistake because terms such as "was found" would be likely indicators for a results phrase. In fact, 'cause' is a key here, since it expresses a biological process or property, but this clue is not salient enough in the unigram model to lead the learner to the correct classification. Allowing extra emphasis to be placed on words which are known in advance to be important to the task may help improve the model. In addition, recognizing the judgmental quality of the adverbial *remarkably* might have indicated to the learner that this sentence was an insight, rather than a strict result.

A number of types of task-specific features employed by [22] were not explored in the present experiments, although many of them may have contributed implicitly through word or lemma n-gram features. There may be room for refinement of these n-gram feature sets. Particular phrases or words may benefit from being highlighted, and leaving some information out may also improve the models. The importance of adverbials is worth investigating, as suggested by the example given above, as is the possibility of creating semantic classes to generalize over words with similar meanings or implications. Among the lexical features experimented with, unigrams and bigrams contributed favorably to models. It remains an avenue worth investigating whether a more limited list of such rhetorical phrases would yield useful features. It also would be worth investigating methods for sub-selecting features in such a way as to maximize the benefits of a given learning algorithm. It may be the case that Naive Bayes and SVMs are best exploited using different approaches to feature selection.

The results for experiments using stopwords shown in Figure 6 suggest that, unlike the case in text classification and IR, the task of ZA benefits from the inclusion of function words. This is likely due to the fact that the task of ZA is connected with the rhetorical organization of a text, in which function words play a role, and furthermore the present analysis is concerned with classifying sentences, as opposed to texts. In the case of sentences, the presence of individual words, even function words, would be likely to have more influence on modeling than they would have on texts. The word *the* for example, occurs in virtually every English language text over a certain length, and is therefore best excluded from unigram models of text. Such words are not a given in every sentence, however, so it is intuitive that they may be able to make contributions to modeling on the sentence level.

6. CONCLUSION

This paper reports the first results on zone analysis experiments in the biomedical domain, using the ZAISA-1 dataset. Several sets of experiments are carried out, including general zone analysis in full texts and specific recognition of results zones within results sections. A variety of intuitive and general-purpose features are explored, yielding a foundation of results upon which to build in the future. The best performing models incorporated a mix of lexical and syntactic features and benefitted considerably by information about location within the text. Incorporating sequentiality into the model by means of a +1-1 window furthermore yielded marked improvement.

Future work will involve further investigation of possible feature sets, expanded to include a variety of task-specific and linguistically motivated features. In particular, specific emphasis will be placed upon words and phrases deemed important to the classification task. Verbs and their modifiers will be given particular focus. Alternate modeling techniques, perhaps including employing multiple learners may also be a promising area of investigation. Research into sub-sentential analysis will also follow from the current work.

7. REFERENCES

[1] G.D. Bader, I. Donaldson, C. Wolting, B.F. Ouellette, T. Pawson, C.W. Hogue. BIND-The Biommolecular Interaction Network Database. *Nucleic Acids Research*, 29:242-245. 2001.

[2] A. Bairoch, R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 200 *Nucleic Acids Research*, 28:302-303. 2000.

[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne. The Protein Data Bank/ *Nucleic Acids Research*, 28:235-242. 2000.

[4] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[5] M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. *ISMB'99*, pp 77-86. 1999.

[6] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[7] S. Dickman. Tough mining; the challenges of searching the scientific literature. *PLoS Biology*, 1(2), pp 144-147. 2003.

[8] K. Humphreys, G. Demetriou, and R. Gaizauskas. Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. In *BSB2000*, pp 502-513. 2000.

[9] T. Joachims. *Learning to Classify Text Using Support Vector Machines*. Kluwer Academic Publishers, 2001.

[10] A. Koike, Y. Kobayashi, and T. Takagi. Kinase pathway database: an integrated protein-kinase and nlp-based protein-interaction resource. *Genome Res*, 17(6A):1231–1243, 2003.

[11] A. Koike and T. Takagi. Prediction of protein-protein interaction sites using support vector machines. *Protein Engineering Design and Selection*, 17(2):165–173, 2004.

[12] L. Lo Conte, S.E. Brenner, T.J.P. Hubbard, C. Chothia, A. Murzin. SCOP database in 2002: Refinements accommodate structural genomics. *Nucleic Acids Research*, 30:264-267. 2002.

[13] Y. Mizuta and N. Collier. An annotation scheme for a rhetorical analysis of biology articles. In *LREC2004*, pp. 1737-1740. 2004.

[14] Y. Mizuta, T. Mullen and N. Collier. Annotation of Biomedical Texts for Zone Analysis. NII Technical Report (NII-2004-007E,ISSN:1346-5597). Oct 2004.

[15] Y. Mizuta, A. Korhonen, T. Mullen and N. Collier. Zone analysis in biology articles as a basis for information extraction. In the *Special Edition on Natural Language Processing in Biomedicine and Its Applications, International Journal of Medical Informatics*. Elsevier. To appear.

[16] S. Novichova, S. Egorov, and N. Darasalia. Medscan, a natural language processing engine for medline abstracts. *Bioinformatics*, 19(13):1699-1706, 2003.

[17] I. Tbahriti, C. Chichester, F Lisacek and P Ruch. Using Argumentation to Retrieve Articles with Similar Citations from MEDLINE. *JNLPBA*, pp 8-14. 2004.

[18] G. Salton and M. J. McGill. The SMART and SIRE Experimental Retrieval Systems. pp.118-155, New York: McGraw-Hill. 1983.

[19] H. Schauer and U. Hahn Phrases as carriers of coherence relations *CogSci 2000—Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pp. 429-434. 2000.

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

[21] P. Tapanainen and T. Järvinen. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing, Washington D.C., Association of Computational Linguistics*, pp 64-71. 1997.

[22] S. Teufel. Arugmentative Zoning: Information Extraction from Scientific Text PhD Thesis. University of Edinburgh. 1999.

[23] S. Teufel and M. Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409-445, 2002.

[24] S. Teufel and H. van Halteren. Agreement in human factoid annotation for summarization evaluation. In *LREC2004*, 2004.

[25] V.N. Vapnik. Statistical Learning Theory. Springer. 1998.

[26] T. Wattarujeekrit, P. Shah and N. Collier PASBio: Predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics* 5:155. 2004.

[27] A. Zanzoni, L. Montecchi, M. Quondam G. Ausiello, M. Helmer-Citterich and G. Cesareni. MINT: A Molecular INTeraction database. *FEBS Lett* 513:135-140. 2002.