Robust Argumentative Zoning for Sensemaking in Scholarly Documents

Simone Teufel[1] and Min-Yen Kan[2]

[1] University of Cambridge Computer Laboratory
sht25@cl.cam.ac.uk
[2] Department of Computer Science, National University of Singapore
kanmy@comp.nus.edu.sg

Abstract. We present an automated approach to classify sentences of scholarly work with respect to their rhetorical function. While previous work that achieves this task of *argumentative zoning* requires richly annotated input, our approach is robust to noise and can process raw text. Even in cases where the input has noise (as it is obtained from optical character recognition or text extraction from PDF files), our robust classifier is largely accurate. We perform an in-depth study of our system both with clean and noisy inputs. We also give preliminary results from *in situ* acceptability testing when the classifier is embedded within a digital library reading environment.

1 Introduction

Even as early as 1984, Cleverdon estimated an annual output of 400,000 papers from the most important journals covering the natural sciences and technology [1]. Today's scholars, even if focusing on a small slice of science that is to become their thesis, need to keep abreast of a large, growing number of scientific developments.

In particular, in the current trend towards interdisciplinarity, researchers will increasingly need to gain an overview of a new field. We call this task *sensemaking*, which is a task that we want to contribute towards. To achieve this goal through the digital library, we need to first generalize some of the needs that researchers must meet. Shum [2] states that what is most interesting to researchers in such a situation is what the main problems and approaches field are. Another question of particular interest is which researchers and groups are connected with which scientific concepts. Knowledge that a scientist acquires over years is a complex network [3]; a system that simply returns an individual publication belies this fact.

Contextual knowledge is needed in order to place and understand the work within the confines of the already existing literature, in all stages of information gathering, *e.g.*, relevance assessment, exploration, reading and utilizing. There is no immediate mechanism in today's digital libraries that addresses this. While most modern digital libraries have keyword search, this ability does little to

R. Bernardi et al. (Eds.): NLP4DL/AT4DL 2009, LNCS 6699, pp. 154–170, 2011.

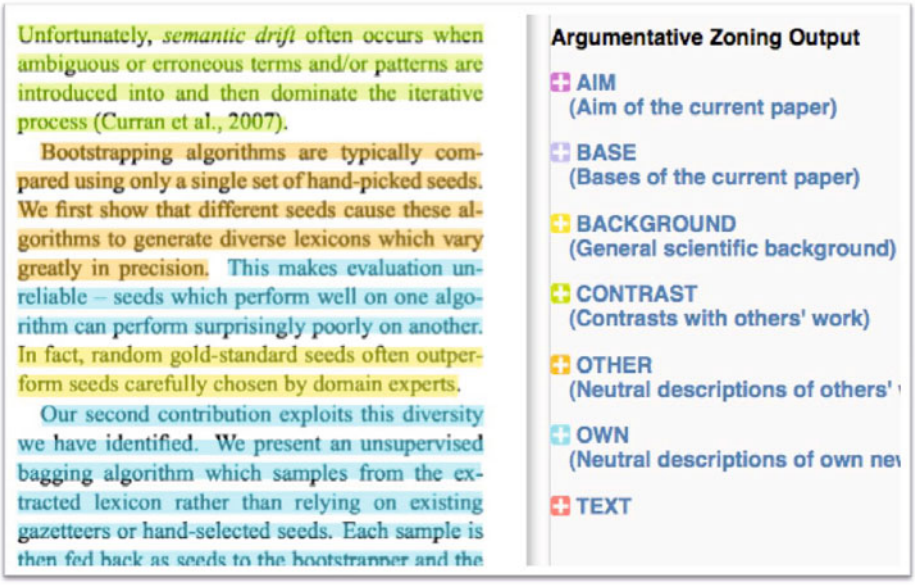

Fig. 1. Argumentative zoning overlaid on a page image from a scholarly article (detail). The sidebar explains the color highlighting of the annotation.

address our challenges [4,5]. What is needed is the provision of assistance so that readers can understand the text.

While as varied as other types of text, scientific discourse is a coherent genre with fixed rhetorical expectations, and with a clear argumentative function. Research articles are biased reports of problem-solving, oriented towards the author's own viewpoint [6]. This fact facilitates the automated analysis of document structure, which largely follows canonical scientific argumentation. Roughly speaking, aims and hypothesis are given first, and are followed by the proof in empirical terms, *e.g.*, the description of an experiment to satisfy the critical and skeptical reader. Particularly important is the embedding of the new work in the research niche, *i.e.*, in relation to already published work. A step towards this sensemaking could be implemented as shown in Figure 1, where a scholarly work is annotated to show which of its sentences discuss the relationship between the work and its contextual literature.

Teufel and Moens [7] introduced *Argumentative Zoning* (AZ), a sentence-based classification of scientific text according to rhetorical status. The AZ classification was designed to be domain-independent and easy for subjects to annotate reliably. In particular, proper AZ annotation highlights how the current work relates to the context of other referenced work in the article.

Given its advantages, it would seem useful to show argumentative zoning alongside an article in a digital library reading environment. However there are

substantial barriers that have thus far prevented the practical, widespread use of AZ. As manual annotation is prohibitively expensive, only an automated system could be considered. However, thus far, automated AZ has only been tried with articles that take rich semantic markup, such as SciXML [8]. Furthermore, to our knowledge, no existing digital library system has fielded a production version of AZ nor shown whether such markup is effective.

Our work in this paper is to address these weaknesses. Specifically, we have created a Robust AZ (RAZ) system that functions over raw English input. We benchmark this system against the original work done previously in Teufel's thesis, which required richly annotated semantic markup, using both clean plain text (extracted from the original richly annotated text) as well as noisy text (extracted directly from the PDF). We have also fielded our classifier within a production digital library system and report preliminary results on the usefulness of such annotation.

2 Argumentative Zoning

Argumentative Zoning (AZ) [7] is an analysis of document structure based on the idea that there are distinct rhetorical moves in scientific papers which together form a scientific argument. An example of a rhetorical move is a goal statement or the criticism of some existing piece of work. The analysis also assumes that rhetorically neutral pieces of text should be classified according to the ownership of the ideas described in the text: are they new contributions (*i.e.*, just being contributed by the authors), statements that nobody in particular lays claim to (*e.g.*, because they are too commonplace), or are they somebody else's (citable) ideas? Another important aspect of the scheme is sentiment, in particular the authors' sentiment towards cited work, as addressed in Nanba and Okumura's work [9].

The categories in the scheme are based on similar rhetorical moves in the literature, *e.g.*, Liddy's Empirical Summary Components [10], Shum's conceptual categories [2], Swales' argumentative moves [6], and Kando's rhetorical categories for information retrieval [11].

AZ is defined as a sentence-based classification according to the following categories (example sentences given in italics; three letter abbreviations in parentheses):

- **Aim (**Aim**)**: Sentences that describe the specific research aims, contributions and conclusions of the current work. *We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts.*
- **Basis (**Bas**)**: Other work that describes tools, theory or findings that the current work uses as a foundation for argument. *The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993).*

- **Background** (BKG): Knowledge that the author feels is generally accepted, not needing argumentative proof or citation. *Methods for automatically classifying words according to their contexts of use have both scientific and practial interest.*
- **Contrast** (CTR): Statements of contrast, comparison, weaknesses of other solutions. These can help identify contradictions or suprising results that differ from established thought. *His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.*
- **Other** (OTH): Other work that is specifically mentioned or cited. Includes work done by the author previously, outside of the current work. *In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events.*
- **Own** (OWN): Sentences that describe the author's own work, method, results, discussion and future work. These sentences comprise the majority of a scholarly document. *More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities p(c|w) for each word w.*
- **Text** (TXT): Sentences that describe the text's internal structure. *We then describe our experimental results in Section 4.*

Teufel *et al.* [12] showed that trained humans are able to produce consistent AZ annotation with acceptable Kappa (κ) [13] of 0.71[1]. Teufel and Moens [7] describe a Naïve Bayes implementation of AZ which is based on 16 sentential features. This model achieves an agreement of $\kappa = 0.45$, whereas Siddharthan and Teufel report $\kappa = 0.48$ for the same data set [14].

3 Related Work

Hachey and Grover [15] present an AZ-based model for the rhetorical classification of legal texts. Their main improvement over Teufel is to use a maximum entropy model, which allows them to use unigrams and bigrams over words as a feature. This improves results considerably. Merity *et al.* [16] use a similar Maximum Entropy approach to AZ which uses unigrams, bigrams and Viterbi search over the category history as its main features. They evalute directly on Teufel and Moens' dataset, and although the evalutation metric used in the paper is not comparable to earlier results (they report weighted accuracy), their classification is more accurate than the earlier results from Teufel and Moens.

A much simpler task than AZ is that of re-introducing rhetorical headlines into structured abstracts in the medical and biological domain [17,18,19]. These typically use structured abstracts to learn a statistical model of what kind of information follows what kinds in abstracts. The models can then be applied to unstructured abstracts in their collection (*e.g.*, only 9% of MEDLINE abstracts are structured).

[1] Kappa values range from 1 (perfect agreement) to -1 (perfect disagreement). A score of 0 indicates no correlation.

4 Method

To accomplish the classification for RAZ, we also turn to maximum entropy (ME) modeling. Like other forms of supervised classification, a ME classifier casts each problem instance as a set of features associated with an appropriate class label. Two key characteristics that differentiate it from other approaches are that the features only take on binary values, and that problem instances are typically chracterized by hundreds of thousands of features. In natural language tasks where word forms are often used as features, the latter characteristic is of utmost importance. Vocabulary sizes in typical English discourse often take a range in the tens of thousands of wordforms.

Each training instance thus can be represented as an n-dimensional feature vector. Even with thousands of training examples, each acting as a constraint on the model, there exist many models that fit the data, as the problem is underconstrained. To select an appropriate model from the multitude possible, the ME classifier seeks out the model where the distribution is most uniform; *i.e.* the model with the maximum entropy.

Finding the unique exact maximum entropy model is usually not possible analytically, but when the feature functions take on an exponential form as in Equation 1, iterative scaling can be used to find a model within an arbitrary ϵ-bound of the exact solution, $\hat{p}$.

$$\hat{p}(y|x) = \frac{1}{Z(x)} exp(\sum_i \lambda_i f_i(x, y)) \tag{1}$$

where $Z(x)$ is the partition function $\sum_y exp(\sum_i \lambda_i f_i(x, y))$, that ensures the $p(\cdot|\cdot)$ values are normalized to actual probabilities. A key consideration of ME is that features for such classifiers do not have to be independent of each other. ME can be implemented to perform feature selection implicitly, so the practitioner is free to introduce a large set of features without much concern with respect to their relevance to the classification task.

During both training and testing, we transform each instance into its vector form: a set of binary valued $f(x, y)$ features. Each feature combines a class label y and a predicate x, as in Equation 2:

$$f(x, y) = \begin{cases} 1 & \text{if } y = \textit{other} \text{ and } x \text{ is the predicate that} \\ & \text{the current sentence contains the word “they”,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

To describe the features for our particular classification task of argumentative zoning, we must describe the classes and predicates. The class labels Y correspond to the set of Teufel's full argumentative zone scheme: {Aim, Bkg, Bas, Ctr, Oth, Own, Txt}. The predicates X fall into different categories of information that we compute from each sentence.

As discussed previously, RAZ takes as input an entire text in plain text (ASCII, UTF-8), processes it to delimit sentences [20] and adds part-of-speech

annotation with part-of-speech tagger [21]. These sentence are fed into the feature computation process, which generates the feature vectors for each sentences. We list these categories below, along with their motivation.

- **Raw tokens:** Individual words in a sentence can be indicative of certain argumentative zone classes. For example, "we" often occurs in a sentence where the authors are describing their own work (OWN). We register each alphanumeric word of the sentence as an individual feature. In addition to the form present in the sentence, we register both lowercased and (English) stemmed forms as different features. Stemming is provided by an implementation of Porter's stemmer [22]. We also capture the word's part-of-speech, to differentiate between different senses of individual words (*e.g.*, "direct" as a adjective or verb). Equation 2 gives an example of a specific word feature.
- **Bigram and Trigram tokens:** Individual words can be ambiguous, and certain word combinations have different meanings and can be strongly indicative of certain classes. For example, "in contrast" strongly signals a contrastive sentence (CTR). We capture contiguous bigram and trigrams from the sentence, and use these as features as well. We create separate bigram and trigram sequences from the raw tokens, as well as from their stemmed form.
- **Cue Words and Phrases:** We look for whether the word contained within a list of 881 known English keywords and 157 cue phrases that may signal a rhetorical move, as defined in Teufel's thesis [23]. She categorized these words and phrases manually in her study of computational linguistics literature. This feature partially overlaps with the previous two classes – "in contrast" and "we" are both listed in these lists – but provide an extra weighting mechanism for the ME classifier to weight the presence of these key terms and phrases more heavily. Some of these cue phrases (about 10%) are actually lexical regular expressions containing part-of-speech constraints which we currently do not handle.
- **Position:** Certain classes of argumentative zones are more prevalent at certain points in the scientific discourse than others. For example, BKG knowledge generally comes in the introduction and surveys of related work. We register the sentence's position in the document, in both absolute and relative terms. We count the number of sentences from the beginning for absolute features, and normalize these versus the number of sentences in the entire document for relative features. Both types of sentence position features are binned at a coarse and fine grained resolution to alleviate problems with data sparsity.
- **Citation Presence:** Citations also strongly indicate certain argumentative classes, such as other work (OTH). Previous studies have differentiated between self-citation (often the basis for the current work; BAS) and citation to others. In RAZ, we built a simple citation presence detector using regular expressions to find standard citation marker patterns. These include numbers in square brackets, tokens that are followed by the suffix "et al." and potential year numbers in parentheses ("[1]", "Wong et al.", "Brown (1988)",

Table 1. Features generated from an example sentence

Sentence (with POS Tagging)	The_DT back-off_JJ model_NN of_IN Katz_NNP (1987)_NNP provides_VBZ a_DT clear_JJ separation_NN between_IN frequent_JJ events,_NN for_IN ...
Tokens	The back off model of katz 1987 provides a clear separation ...
	STEMSthe STEMSback-off STEMSmodel STEMSof STEMSkatz STEMS(1987 ...
Bigrams / Trigrams	The_DT_back-off_JJ back-off_JJ_model_NN model_NN_of_IN of_IN_Katz_NNP ...
	The_DT_back-off_JJ_model_NN back-off_JJ_model_NN_of_IN ...
Cue Phrases	CPWORK CPPOS
Sentence Position	REL_POSITION2 ABS_POSITION1 REL_POSITION2_1 ABS_POSITION25
Citation Presence	CITEyear CITEcount1
Sentence Length	SENTLENGTH3
Title Overlap	*(N/A)*
Agent	AGENTmodel_nn
Verb tense	VERBprovides_vbz VERBTENSEvbz

respectively). Our citation detector is quite simple, aiming for a balance of precision and recall while maintaining efficiency.

- **Sentence Length:** Longer sentences can correlate to detailed discussion and data analysis. We measure the length of a sentence in ten word units as a feature.
- **Title Overlap:** If a sentence's words overlap with the title, there is a higher probability that it elaborates on the theme of the article (*e.g.*, OWN). We treat the first 100 words of the article as a "title" and identify which words in a candidate sentence overlap with these title words. We use this span because in the raw input text, we have no explicit way to capture the title, so we use this approximation.
- **Agent:** Syntactic information can further discriminate the role of certain words. The token "we" can be the agent of a sentence (*e.g.*, "We performed...") or can be the patient receiving an action ("...is different from what we measure"). Given the part-of-speech input, we use a set of simple heuristics to locate the agent of the sentence, and encode this as an individual feature.
- **Verb tense:** Similarly, verb identity and tense can also signal particular argumentative classes. Sentences in past tense can disclose past work (OTH), for example. We use the part-of-speech information to locate the main verb in the sentence, using a set of heuristics, and create features for its identity and tense.

Finally we feed the feature vectors to the maximum entropy software[2]) to generate models in training, or to label new unseen sentences in testing.

Table 1 illustrates a concrete example of the different features that are computed, given a sample sentence.

5 Evaluation

Our formal evaluation tests our RAZ system in with both perfect input (with correct splitting of sentences) as well as realistic, noisy input (using automatic

[2] We use Le Zhang's toolkit, available at: http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

sentence splitting and part of speech tagging done programatically). We benchmark RAZ against the original AZ system devised by Teufel [23], which used richly annotated SciXML as input.

In contrast, our RAZ system handles impoverished input of plain text that has been extracted from PDF files. It is well known to the community that extracting text from PDF files (especially legacy PDF files) can be problematic, due to accents, ligatures (*e.g.*, "fi" combined into a single glyph) and font substitutions.

In our formal evaluation, we wish to answer the following questions to better understand RAZ.

Question A. How much does argumentative zoning recognition decrease if we use clean text instead of the "perfect" semantically rich markup provided by SciXML?

Question B. How much does argumentative zoning recognition decrease if we use noisy text instead of clean text?

Question C. What performance is achieved in using the different sets of features? How important is each feature class towards achieving the maximum classification accuracy?

Question D. What types of errors commonly occur in the best performing classifier?

5.1 Corpus

We obtained the 74 gold standard files from [23], which represent open-access computational linguistics research articles contributed to the arXiV digital library from a period of 1994 to 1996. These have "perfect" XML structure and "perfect" human AZ annotation; we call this set "Dataset P". The SciXML markup used in Dataset-P provides annotations of correct paragraph and sentence boundaries, topic changes, hierarchical logical structure (including headers), equations, citations (differentiating self-citations from others), and citation function.

We then further stripped Dataset-P of the rich markup XML, to reveal "clean" text, which is however not perfect because it lacks the important structural markup. We call this set of "Dataset C" (for "clean").

Finally, we located the original corresponding 74 source .PDF files from arXiv. By programmatically extracting the text from the PDF files, we obtained a final

Table 2. Dataset descriptions

	Perfect (Dataset P)	Clean (Dataset C)	Noisy (Dataset N)
Text obtained via	Manual Entry	Manual Entry	Automatic Extraction from PDF
Structural markup (XML)	Hand-annotated	Absent	Absent
Paragraph and Sentence Boundaries	Hand-corrected	Hand-corrected	Automatic
POS Tags	Hand-corrected	Hand-corrected	Automatic

"noisy" dataset, complete with all imperfections that come with such a method – incorrectly extracted words, hypenation, font substitutions and occasional column flow problems. We call this set of papers "Dataset N" (for "noisy").

The performance of Dataset N against the gold standard Dataset P measures the system's errors. Dataset C measures the portion of system errors that are attributable to differences in modeling (Q1). The difference between Datasets C and N, however, quantifies the system's robustness (Q2); *i.e.*, how much decrease in performance is due to the textual noise in the PDF texts, as opposed to the loss of structure information. The perfect AZ data set has both clean text and structure information, but the implementation of all AZ features is only possible with structural information.

5.2 Noisy Evaluation – Questions A and B

A heuristic alignment of sentences in Dataset P with Datasets N and C is necessary, as the automatic creation of Dataset N from the PDF incurs errors in detecting sentence boundaries, in detecting non-running text (such as titles, authors, headers, footnotes, etc), and might incorporate non-running text partially into "sentences". Our implementation uses edit-distance, by calculating the ratio of the longest common substring shared between two potentially aligned sentences, to their average length [24]. Matches are accepted if a threshold (currently 0.65) is exceeded; heuristic search attempts to maintain relative sentence ordering, but can jump over up to 30 sentences, as the PDF conversion is often unable to exclude non-running text which occurs as tables or figures.

The alignment reveals a precision and recall of aligning Dataset P with Dataset N of 9.71% and 15.88%, which is very low. This number might underestimate the real precision and recall, which we believe to be in the range of 70%, but latest measurements were not possible due to time limits. The numbers are also lowered by the fact that Dataset P has an idiosyncratic marking of sentences containing equations. This results in Dataset C, although the text is entirely clean, also does not reach 100% precision and recall on alignment: Precision of aligning Dataset P with Dataset C is 0.97, and recall is 0.99.

Once sentences are aligned, normal agreement figures can be reported for both Dataset N and C. We use 2-fold cross-validation.

Results and Discussion. Table 3 shows the results of the comparison to the 74 gold standard files. Agreement is measured using κ, which corrects for chance agreement [25,26,27]. We also report accuracy $P(A)$, chance agreement $P(E)$, number of items (N) and 95% confidence interval for κ.[3]

Table 3 answers Questions A and B. On first inspection, RAZ in both its clean and noisy incarnations, fares significantly worse than the previously reported AZ system that uses "perfect" data. However, one should note that the ceiling we compare against (AZ with 16 features as reported in [14]) is not directly comparable, as 6 additional files are used in their case.

[3] Reporting *kappa* with a confidence interval is one of the recommendations brought forward in [28].

Table 3. Agreement with gold standard for RAZ with Noisy and Clean input data, in comparison to AZ with Perfect input from [14]

Proc.	Dataset; # Files	X-val Folds	κ	N	$P(A)$	$P(E)$
RAZ	N (74)	2	0.23±0.016	8,494	0.63	0.53
RAZ	C (74)	2	0.28±0.014	11,732	0.68	0.56
AZ	P (80)	10	0.48±0.014	12,464	0.76	0.54

The good news is that RAZ fares respectably on Dataset N when compared to Dataset C, answering Question B, and validating our claim of robustness. When interpreting this gap, one also needs to consider that on Dataset N, performance can only be evaluated on *aligned* sentences, whereas RAZ on Dataset C is evaluated on practically all sentences (alignment is trivial, as the RAZ pipeline was given texts from original gold standard corpus[4]. Thus, Dataset N has far fewer sentences than Dataset C and would be perceived by a human user as obviously inferior, due to this fact.

The RAZ results at $\kappa = 0.23$ (Dataset N) and $\kappa = 0.28$ (Dataset C) are respectable considering how little information the classifier has at its disposal, in comparison to full AZ, where structural information, syntactic information, and full regular expressions for meta-discourse can be exploited. One should also take into consideration that RAZ classification is immediately and practically usable, in contrast to any other AZ implementation we know of: it is robust, can be performed on practically any scientific text available on the web and it produces its classification in real-time (about 10 milliseconds per sentence on a standard desktop PC; equivalent to 3-5 seconds per conference paper).

5.3 Clean Evaluation – Questions C and D

In our clean evaluation, we examine the performance over Dataset C, the dataset used to train the final, deployed RAZ classifier. Here, we wish to assess the usefulness of individual feature classes towards overall classification performance. We used 10-fold stratified cross validation to assess our ME classification model performance with the differing feature classes introduced previously. Table 4 gives the raw accuracy, macro averaged precision, recall and F_1 performance levels for these different combinations.

In addition to this macro-level analysis, we also wish to assess the performance of individual AZ categories. As such, we carried out a more detailed error analysis. Table 5 gives the full confusion matrix among the classifier's decision using the full model that utilizes all features.

Results and Discussion. Question C of our evaluation is answered by the data in Table 4. Surprisingly, performance peaks (when measured by macro F_1) when we use all of the features except the bigrams and trigrams. We believe this is caused by the sparsity of data that comes from this feature, causing minority AZ classes to suffer. As there is some redundancy between the word (unigram)

[4] Of course the system was never tested on any text it was trained on.

Table 4. Feature ablation test performance, averaged over stratified 10-fold cross valiation. Precision, Recall and F_1 are macro averaged over all 7 AZ categories.

Feature Classes	Accuracy	Precision	Recall	F_1
All	66.8%	47.8%	37.6%	.4142
All features except one				
All − Words	67.4%	45.5%	33.4%	.3739
All − 2,3 grams	66.4%	46.8%	40.9%	.4339
All − Title	66.7%	47.1%	38.0%	.4151
All − Sent Position	65.3%	45.6%	35.5%	.3908
All − Cue Phrases	66.6%	47.2%	36.2%	.4019
All − Cite	67.1%	49.3%	36.9%	.4121
All − Sent Length	66.8%	46.7%	37.8%	.4116
All − Agent	66.6%	47.1%	37.4%	.4102
All − Verb	67.4%	48.7%	37.3%	.4133
Single features				
Words	59.7%	36.9%	34.4%	.3553
2,3 grams	59.0%	34.9%	31.2%	.3268
Title	–	–	–	–
Sent Position	66.9%	18.2%	20.6%	.1810
Cue Phrases	22.4%	23.8%	8.8%	.1229
Cite	68.3%	16.1%	17.2%	.1604
Sent Length	66.8%	9.5%	14.2%	.1145
Agent	52.5%	38.0%	22.7%	.2686
Verb	37.9%	23.0%	10.7%	.1332

Table 5. Confusion matrix for the RAZ classifier using all feature classes for our clean evaluation. Gold standard answers in rows, RAZ automatic classification in columns. Bolded figures are correct classification instances. The *Undefined* (Un.) class is used for text that is not body text (*e.g.*, section headers, page numbers).

n=12898	Aim	Bas	Bkg	Ctr	Oth	Own	Txt	Un.	# of instances	Precision	Recall	F_1
Aim	**138**	5	14	10	7	50	4	1	229 (1.77%)	60%	44%	51%
Bas	10	**45**	5	10	39	45	1		155 (1.20%)	29%	18%	22%
Bkg	10	4	**157**	46	105	167	2	2	493 (3.82%)	32%	20%	24%
Ctr	3	6	40	**86**	52	111	4		302 (2.34%)	28%	14%	19%
Oth	16	60	148	102	**559**	695	14	4	1598 (12.38%)	35%	28%	31%
Own	131	119	419	342	1253	**7526**	85	14	9889 (76.67%)	76%	87%	81%
Txt	5	3	2	1	4	26	**117**		158 (1.22%)	74%	52%	61%
Un.			4	4	1		3	**62**	74 (0.5%)	–	–	–

and the bi-/tri-gram features, omitting either one does not cause much change in the model.

It is more obvious which features are most significant to the ME models when only a single feature class was used. All single feature models underperform the combined classifiers significantly. While some simple models (*e.g.*, Sentence Position, Citation, Sentence Length) are as accurate as the combined classifiers on a per-instance basis, their F_1 scores are dismal (≈11–.18), showing that they

mostly just classify all sentences as OWN, the majority class. Our tests show that the battery of features is robust on its own (from our All – single feature tests), and that no single feature performs well outright (from our single feature tests).

To answer Question D, we turn to the detailed analysis of the confusion matrix of the full classifier (Table 5). Focusing first on the number of instances of each class, we notice right away the problem of skewed input in the dataset – almost 90% of the ground truth belongs to just two classes: OTH and OWN. This skewed input certainly makes the recognition of the minority classes difficult, as only a modicum of training data is available for these classes.

Among the remaining 5 minority classes, textual structure (TXT) and aims (AIM) are relative easy to identify, likely due to the presence of key words (*e.g.*, "propose", "Section") and common positions (following other TXT, or at the beginning of the paper). Background, Basis and Other are also commonly mistaken for each other, due to their similarity in wording. This is also a common mistake for people to make as well – in some sense, all three of these classes describe contextual information needed to understand the author's own claims, but differ in the nuances of attribution. We believe being able to attribute personal names and citations to either the paper's authors (self-citation) or to others would help to improve these classes' recognition. Finally, the contrast class CTR is the most difficult to classify, with a meagre .19 F_1. Contrasts are sometimes built over multiple sentences and are not always signaled explicitly by discourse cues, contributing to false negatives. On the other hand, some strong lexical cues for contrast are also used in other ways (*e.g.*,"We were able to detect the objects however small they appear in the video dataset"), leading to false positives.

6 Deployment

We believe that argumentative zoning is useful in obtaining an overview of a document's purpose, structure of argumentative and relationship to other documents. To test this theory, we must integrate the AZ classifer within a digital library reading interface, where the reader can view AZ annotations directly on the document. For this purpose, we retrained the RAZ classifier over the full training dataset, and incorporated it into ForeCite [29], a digital library that has a web-based reading environment that can display arbitrary, word-span based annotations, as shown in earlier in Figure 1 (which is actually a detail of the screenshot of the system), and in Figure 2.

The interface overlays a transparent colored layer over each sentence in the document, where the color is determined by the RAZ classifier. The interface allows the reader to see the AZ annotation of a sentence in the context of other sentences (in the reader window, left panel), as well as jump to other parts of the document, grouped by AZ classes (right panel). The AZ panel features a collapsable hierarchical interface that allows quick access to the text and location of sentences of a particular AZ class.

The careful reader will note that RAZ annotation is omitted from the bullet points and headers in 2. The ForeCite framework automatically determines (with

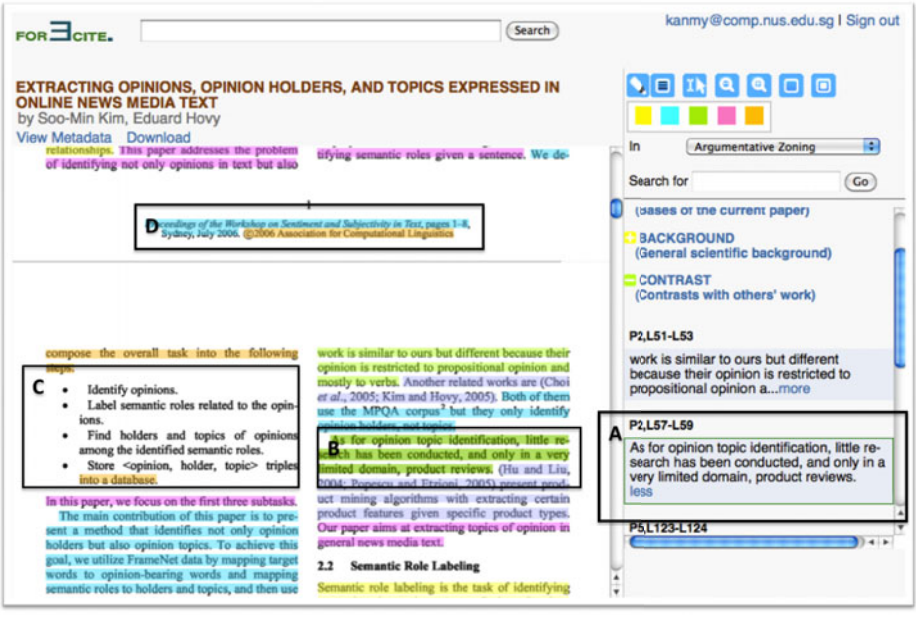

Fig. 2. RAZ annotations in the ForeCiteReader environment. The reader has clicked on a particular Contrast sentence in the sidebar (Inset A), which has been automatically highlighted in the reader (Inset B, slightly darker green than other CTR sentences). Some lines have been misidentified by the environment as non-body text and vice versa (Insets C and D).

some noise) which text in the document is body text, where sentences start and end, and passes only these body text sentences to the RAZ classifier.

6.1 In Situ Usability

To assess whether RAZ does help in sensemaking, we have carried out a preliminary indicative study with the ForeCite reading interface. The task in the experiment was to answer four central questions about the paper, and then to critique the interface and annotation shown.

Four graduate research students in information systems were asked to skim two computational linguistics 8-page conference papers (not from the training data), for which they had little previous background. However, they had general background in reading conference papers. They were shown one document in the ForeCiteReader plain reading interface, and one displaying RAZ annotation. Each subject experienced a different ordering of the papers and of the interfaces, to mitigate order and learning effects.

The students were given a copy of the task instructions, while the interviewer verbally went over the instructions. Before answering the questions and providing feedback, the students were to first skim the documents. The entire interview

took about 30 minutes per subject. The four questions that were given are replicated below, paraphrased for length considerations:

Q1: Please name the central contribution(s) of the work.
Q2: Name two related works in this paper and describe their relationship to this paper.
Q3: Identify any datasets that were used in evaluation and its origin.
Q4: (*one variant shown*) Let's suppose that another paper cites this paper as: "*[This paper] describes the first approach to apply co-training in a bilingual setting, that is with a pair of languages.*" Can you identify whether this citation claim is a valid or not?

The students were informed that they would be timed, but that a longer or shorter time to task completion would not affect them in any way. In designing the questionnaire, we hypothesized that Q1 and Q2 could be addressed by using the AZ markup, specifically by the AIM for Q1 and OTH, CTR and BAS classes, for Q2. Q3 was inserted as a control, as AZ does not specifically indicate sentences that describe datasets (as the AZ scheme is general and does not presume experimental validation). Q4 was tailored to each of the two articles, and asks whether a claim in another (hypothetical) paper citing the target paper could be validated.

We emphasize that our study is indicative and not designed to be summative or statistically accurate, since the sample size is small. With this in mind, we discuss salient observations from the survey with respect to AZ.

Time to Task Completion. AZ did not have a measureable effect on task time completion. Using AZ required the subject to experiment with the interface and also required subjects to shift attention (the left vs. right of the interface) and to change task (reading vs. focused navigation). The sentence previews in the AZ sidebar alleviated this somewhat, but when context was needed to interpret the sentence, subjects had to return to the reading panel to verify evidence.

AZ Effectiveness. In both subjective opinion and interviewer observation, AZ had a positive effect in locating answers to Q2 when used by the subjects. For Q1, it did not help as both papers indicated the goal within the abstract or introduction, and most of the subjects started off reading. However, one subjects did use the AIM class to read off the contributions of the paper as listed throughout different sections of the paper.

Annotation Noise. While our current AZ classifier performs only at a mediocre level (.41 macro F_1), differentiating the minority classes (*i.e.*, AIM, BAS, CTR, OTH) from the OWN majority helped to identify candidate sentences that might contain answers. Subject commented that the annotations were largely correct for key sentences and that errors in the automatic segmentation of the body text and sentence delimitation were larger barriers than the AZ classification itself (as seen in Figure 2).

Interaction with Reading. Subjects universally complained that the AZ coloring detracted from their reading experience, as it decreased the contrast

of the text. Subjects suggested that the interface should be loaded plain but that spans could be colored on demand from the right hand side AZ sidebar.

Unknown Terminology. While unknown terminology kept subjects from deep understanding of the goal, they were generally able to recognize sentences that describe answers to the questions. Two subjects stated that an extension of AZ could help identify definitions.

Other Extensions of AZ. Two of the four subjects suggested that AZ be extended to work to extract key facts, such as the identity of datasets and specific tools or methodologies used. They also mentioned that contributions (rather than aims) and experimental results could be part of the AZ schema.

The pilot evaluation hints that RAZ can be a useful part of a holistic sense-making interface. Our current RAZ system certainly assists the reader, along with the authors' own careful text structuring, in interpreting the major points in a work as well as its contextual place among referenced documents. While the reading interface should place reading functionality first, AZ (and possibly other) annotations should be called on demand and displayed in an unobtrusive manner.

We are currently revising the integrated system to account for the feedback. Our current work can be categorized along two fronts. The first front is in improving the interface, such that reading ease is maximized. Standard, digital reading affordances (locating a section, page, or finding instances of individual words) need to be supported. Parallel work within the ForeCite digital library project has achieved this goal of the re-discovery of logical sections from scientific documents (both modern ones born digitally as well as legacy documents that are only represented by scanned images) [30].

Second, as our subjects have commented, AZ highlighting should be done on demand by the user and only for a particular AZ class. We have modified the display framework to account for this feedback. Own sentences in particular are not helpful to identify (as this is the majority of the paper), so are now omitted from the interface entirely. The second front is to extend information extraction and classification further into the document content. Our longer term plans are to extract definitions of terms, identify the semantic categories of pertinent keyphrases as methods, systems, tools, or other domain-specific constructs. These may further aid the understanding of the document.

7 Conclusion

Abstracts have been acknowledged as the author's view of the importance of their own work. Recently, the community has acknowledged that sentences that cite a paper describe the community's view of the importance of a paper [31,32]. We claim that the document itself has its own voice about what is important. The discourse and argumentative structure in a well-written paper also direct a reader to its important aspects within the reading context.

We have captured this notion of argumentative zoning (AZ) in an implemented classifier and described the textual features it uses to render its judgment.

To our knowledge, we describe the first robust AZ system (RAZ) that is able to perform such classification on noisy inputs that come from PDF text extraction, as well as the relatively clean output of optical character recognition.

Our work also represents the first system that has been integrated into a production digital library system, ForeCite. Our preliminary *in situ* study indicates that robust AZ can be a helpful source of evidence in sensemaking for understanding the contributions and context of the individual scholarly paper.

Acknowledgments

The second author would like to acknowledge the help of the graduate student volunteers at UC Irvine for their help in evaluating the RAZ interface.

References

1. Cleverdon, C.W.: Optimizing convenient online access to bibliographic databases. Information Services and Use 4, 37–47 (1984)
2. Shum, S.B.: Evolving the web for scientific knowledge: First steps towards an "HCI knowledge web". Interfaces, British HCI Group Magazine 39, 16–21 (1998)
3. Bazerman, C.: Physicists reading physics, schema-laden purposes and purpose-laden schema. Written Communication 2(1), 3–23 (1985)
4. Kircz, J.G.: The rhetorical structure of scientific articles: The case for argumentational analysis in information retrieval. Journal of Documentation 47(4), 354–372 (1991)
5. Ingwersen, P.: Cognitive perspectives of information retrieval interaction: Elements of a cognitive ir theory. Journal of Documentation 52, 3–50 (1996)
6. Swales, J.: Research articles in English. In: Genre Analysis: English in Academic and Research Settings, ch. 7, pp. 110–176. Cambridge University Press, Cambridge (1990)
7. Teufel, S., Moens, M.: Summarising scientific articles — experiments with relevance and rhetorical status. Computational Linguistics 28(4), 409–446 (2002)
8. Copestake, A., Corbett, P.T., Murray-Rust, P., Rupp, C.J., Siddharthan, A., Teufel, S., Waldron, B.: An architecture for language technology for processing scientific texts. In: UK e-Science All Hands Meeting (2006)
9. Nanba, H., Okumura, M.: Towards multi-paper summarization using reference information. In: Proceedings of IJCAI 1999, pp. 926–931 (1999)
10. Liddy, E.D.: The discourse-level structure of empirical abstracts: An exploratory study. Information Processing and Management 27(1), 55–81 (1991)
11. Kando, N.: Text-level structure of research papers: Implications for text-based information processing systems. In: Proceedings of BCS-IRSG Colloquium, pp. 68–81 (1997)
12. Teufel, S., Carletta, J., Moens, M.: An annotation scheme for discourse-level argumentation in research articles. In: Proceedings of European ACL (EACL 1999), Bergen, Norway, pp. 110–117 (1999)
13. Siegel, S., Castellan, N.J.J.: Nonparametric Statistics for the Behavioral Sciences, 2nd edn. McGraw-Hill, Berkeley (1988)

14. Siddharthan, A., Teufel, S.: Whose idea was this, and why does it matter? attributing scientific work to citations. In: Proceedings of the North American chapter of the Association of Computational Linguistics, NAACL 2007 (2007)
15. Hachey, B., Grover, C.: Extractive summarisation of legal texts. Artificial Intelligence and Law: Special Issue on E-government 14(4), 305–345 (2006)
16. Merity, S., Murphy, T., Curran, J.R.: Accurate argumentative zoning with maximum entropy models. In: Proceedings of ACL-IJCNLP 2009 Workshop on text and citation analysis for scholarly digital libraries (NLPIR4DL), Singapore, pp. 19–26 (2009)
17. McKnight, L., Arinivasan, P.: Categorization of sentence types in medical abstracts. In: AMIA 2003 Symposium Proceedings, pp. 440–444 (2003)
18. Lin, J., Karakos, D., Demner-Fushman, D., Khudanpur, S.: Generative content models for structural analysis of medical abstracts. In: Proceedings of the HLT/NAACL 2006 Workshop on Biomedical Natural Language Processing (BIONLP 2006), New York City, USA, pp. 65–72 (2006)
19. Hirohata, K., Okazaki, N., Ananiadou, S., Ishizuka, M.: Identifying sections in scientific abstracts using conditional random fields. In: Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008), Hyderabad, India, pp. 381–388 (2008) ACL Anthology Ref. I08-1050
20. Reynar, J.C., Ratnaparkhi, A.: A maximum entropy approach to identifying sentence boundaries. In: Proceedings of the Firth Conference on Applied Natural Language Processing, pp. 803–806 (1997)
21. Ratnaparkhi, A.: A maximum entropy model for part-of-speech tagging. In: Empirical Methods for Natural Language Processing. Association for Computational Linguistics, New Jersey (1996)
22. Porter, M.F.: An algorithm for suffix stripping. Program (3), 130–137 (1980)
23. Teufel, S.: Argumentative Zoning: Information Extraction from Scientific Text. PhD thesis, School of Cognitive Science, University of Edinburgh, Edinburgh, UK (2000)
24. Hirschberg, D.S.: A linear space algorithm for computing maximal common subsequences. Communications of the ACM 18(6), 341–343 (1975)
25. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76, 378–381 (1971)
26. Carletta, J.: Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics 22(2), 249–254 (1996)
27. Krippendorff, K.: Content Analysis: An Introduction to its Methodology, 2nd edn. Sage Publications, Beverly Hills (2004)
28. Krenn, B., Evert, S., Zinsmeister, H.: Determining intercoder agreement for a collocation identification task. In: Proceedings of Konvens 2004 (2004)
29. Nguyen, T.D., Kan, M.Y., Dang, D.T., Hänse, M., Hong, C.H.A., Luong, M.T., Gozali, J.P., Sugiyama, K., Tan, Y.F.: ForeCite: towards a reader-centric scholarly digital library. Under Review (2010)
30. Luong, M.T., Nguyen, T.D., Kan, M.Y.: Logical structure recovery in scholarly articles with rich document features. International Journal of Digital Library Systems (2011)
31. Nakov, P., Schwarz, A., Hearst, M.: Citances: Citation sentences for semantic analysis of bioscience text. In: SIGIR 2004 Workshop on Search and Discovery in Bioinformatics (2004)
32. Qazvinian, V., Radev, D.R.: Scientific paper summarization using citation summary networks. In: Proceedings of COLING 2008, Manchester, UK (2008)