Towards Automatic Classification of Discourse Elements in Essays

Jill Burstein
ETS Technologies
MS 18E
Princeton, NJ 08541
USA
Jburstein@
etstechnologies.com

Daniel Marcu
ISI/USC
4676 Admiralty
Way
Marina del Rey,
CA, USA
Marcu@isi.edu

Slava Andreyev
ETS Technologies
MS 18E
Princeton, NJ 08541
USA
sandreyev@
etstechnologies.com

Martin Chodorow
Hunter College, The
City University of
New York
New York, NY USA
Martin.chodorow@
hunter.cuny.edu

Abstract

Educators are interested in essay evaluation systems that include feedback about writing features that can facilitate the essay revision process. For instance, if the thesis statement of a student's essay could be automatically identified, the student could then use this information to reflect on the thesis statement with regard to its quality, and its relationship to other discourse elements in the essay. Using a relatively small corpus of manually annotated data, we use Bayesian classification to identify thesis statements. This method yields results that are much closer to human performance than the results produced by two baseline systems.

1 Introduction

Automated essay scoring technology can achieve agreement with a single human judge that is comparable to agreement between two single human judges (Burstein, et al 1998; Foltz, et al 1998; Larkey, 1998; and Page and Peterson, 1995). Unfortunately, providing students with just a score (grade) is insufficient for instruction. To help students improve their writing skills, writing evaluation systems need to provide feedback that is specific to each individual's writing and that is applicable to essay revision.

The factors that contribute to improvement of student writing include refined sentence structure, variety of appropriate word usage, and organizational structure. The improvement of organizational structure is believed to be critical in the essay revision process toward overall improvement of essay quality. Therefore, it would be desirable to have a system that could indicate as feedback to students, the discourse elements in their essays. Such a system could present to students a guided list of questions to consider about the quality of the discourse.

For instance, it has been suggested by writing experts that if the *thesis statement*[1] of a student's essay could be automatically provided, the student could then use this information to reflect on the thesis statement and its quality. In addition, such an instructional application could utilize the thesis statement to discuss other types of discourse elements in the essay, such as the relationship between the *thesis statement* and the *conclusion*, and the connection between the *thesis statement* and the *main points* in the essay. In the teaching of writing, in order to facilitate the revision process, students are often presented with 'Revision Checklists.' A revision checklist is a list of questions posed to the student to help the student reflect on the quality of his or her writing. Such a list might pose questions such as:

a) Is the intention of my thesis statement clear?

[1] A *thesis statement* is generally defined as the sentence that explicitly identifies the purpose of the paper or previews its main ideas. See the Literacy Education On-line (LEO) site at http://leo.stcloudstate.edu.

(Annotator 1) "In my opinion student should do what they want to do because they feel everything and they can't have anythig they feel because they probably feel to do just because other people do it not they want it.

(Annotator 2) I think doing what students want is good for them*. I sure they want to achieve in the highest place but most of the student give up. They they don't get what they want. To get what they want, they have to be so strong and take the lesson from their parents Even take a risk, go to the library, and study hard by doing different thing.*

Some student they do not get what they want because of their family. Their family might be careless about their children so this kind of student who does not get support, loving from their family might not get what he wants. He just going to do what he feels right away.

So student need a support from their family they has to learn from them and from their background. I learn from my background I will be the first generation who is going to gradguate from university that is what I want."

Figure 1: Sample student essay with human annotations of thesis statements.

b) Does my thesis statement respond directly to the essay question?
c) Are the main points in my essay clearly stated?
d) Do the main points in my essay relate to my original thesis statement?

If these questions are expressed in general terms, they are of little help; to be useful, they need to be grounded and need to refer explicitly to the essays students write (Scardamalia and Bereiter, 1985; White 1994). The ability to automatically identify and present to students the discourse elements in their essays can help them focus and reflect on the critical discourse structure of the essays. In addition, the ability for the application to indicate to the student that a discourse element could not be located, perhaps due to the 'lack of clarity' of this element, could also be helpful. Assuming that such a capability was reliable, this would force the writer to think about the clarity of an intended discourse element, such as a thesis statement.

Using a relatively small corpus of essay data where thesis statements have been manually annotated, we built a Bayesian classifier using the following features: sentence position; words commonly used in thesis statements; and discourse features, based on Rhetorical Structure Theory (RST) parses (Mann and Thompson, 1988 and Marcu, 2000). Our results indicate that this classification technique may be used toward automatic identification of thesis statements in essays. Furthermore, we show that this method generalizes across essay topics.

2 What Are Thesis Statements?

A *thesis statement* is defined as the sentence that explicitly identifies the purpose of the paper or previews its main ideas (see footnote 1). This definition seems straightforward enough, and would lead one to believe that even for people to identify the thesis statement in an essay would be clear-cut. However, the essay in Figure 1 is a common example of the kind of first-draft writing that our system has to handle. Figure 1 shows a student response to the essay question:

> *Often in life we experience a conflict in choosing between something we "want" to do and something we feel we "should" do. In your opinion, are there any circumstances in which it is better for people to do what they "want" to do rather than what they feel they "should" do? Support your position with evidence from your own experience or your observations of other people.*

The writing in Figure 1 illustrates one kind of challenge in automatic identification of discourse elements, such as thesis statements. In this case, the two human annotators independently chose different text as the thesis statement (the two texts highlighted in bold and italics in Figure 1). In this kind of first-draft writing, it is not uncommon for writers to repeat ideas, or express more than one general opinion about the topic, resulting in text that seems to contain multiple thesis statements.

Before building a system that automatically identifies thesis statements in essays, we wanted to determine whether the task was well-defined. In collaboration with two writing experts, a simple

discourse-based annotation protocol was developed to manually annotate discourse elements in essays for a single essay topic. This was the initial attempt to annotate essay data using discourse elements generally associated with essay structure, such as *thesis statement, concluding statement,* and *topic sentences of the essay's main ideas*. The writing experts defined the characteristics of the discourse labels. These experts then annotated 100 essay responses to one English Proficiency Test (EPT) question, called Topic B, using a PC-based interface implemented in Java.

We computed the agreement between the two human annotators using the kappa coefficient (Siegel and Castellan, 1988), a statistic used extensively in previous empirical studies of discourse. The kappa statistic measures pairwise agreement among a set of coders who make categorial judgments, correcting for chance expected agreement. The kappa agreement between the two annotators with respect to the thesis statement labels was 0.733 (N=2391, where 2391 represents the total number of sentences across all annotated essay responses). This shows high agreement based on research in content analysis (Krippendorff, 1980) that suggests that values of kappa higher than 0.8 reflect very high agreement and values higher than 0.6 reflect good agreement. The corresponding *z* statistic was 27.1, which reflects a confidence level that is much higher than 0.01, for which the corresponding *z* value is 2.32 (Siegel and Castellan, 1988).

In the early stages of our project, it was suggested to us that thesis statements reflect the most important sentences in essays. In terms of summarization, these sentences would represent indicative, generic summaries (Mani and Maybury, 1999; Marcu, 2000). To test this hypothesis (and estimate the adequacy of using summarization technology for identifying thesis statements), we carried out an additional experiment. The same annotation tool was used with two different human judges, who were asked this time to identify the most important sentence of each essay. The agreement between human judges on the task of identifying summary sentences was significantly lower: the kappa was 0.603 (N=2391). Tables 1a and 1b summarize the results of the annotation experiments.

Table 1a shows the degree of agreement between human judges on the task of identifying thesis statements and generic summary sentences. The agreement figures are given using the kappa statistic and the relative precision (P), recall (R), and F-values (F), which reflect the ability of one judge to identify the sentences labeled as thesis statements or summary sentences by the other judge. The results in Table 1a show that the task of thesis statement identification is much better defined than the task of identifying important summary sentences. In addition, Table 1b indicates that there is very little overlap between thesis and generic summary sentences: just 6% of the summary sentences were labeled by human judges as thesis statement sentences. This strongly suggests that there are critical differences between thesis statements and summary sentences, at least in first-draft essay writing. It is possible that thesis statements reflect an intentional facet (Grosz and Sidner, 1986) of language, while summary sentences reflect a semantic one (Martin, 1992). More detailed experiments need to be carried out though before proper conclusions can be derived.

Table 1a: Agreement between human judges on thesis and summary sentence identification.

Metric	Thesis Statements	Summary Sentences
Kappa	0.733	0.603
P (1 vs. 2)	0.73	0.44
R (1 vs. 2)	0.69	0.60
F (1 vs. 2)	0.71	0.51

Table 1b: Percent overlap between human labeled thesis statements and summary sentences.

	Thesis statements vs. Summary sentences
Percent Overlap	0.06

The results in Table 1a provide an estimate for an upper bound of a thesis statement identification algorithm. If one can build an automatic classifier that identifies thesis statements at recall and precision levels as high as 70%, the performance of such a classifier will be indistinguishable from the performance of humans.

3 A Bayesian Classifier for Identifying Thesis Statements

3.1 Description of the Approach

We initially built a Bayesian classifier for thesis statements using essay responses to one English Proficiency Test (EPT) test question: Topic B.

McCallum and Nigam (1998) discuss two probabilistic models for text classification that can be used to train Bayesian independence classifiers. They describe the multinominal model as being the more traditional approach for statistical language modeling (especially in speech recognition applications), where a document is represented by a set of word occurrences, and where probability estimates reflect the number of word occurrences in a document. In using the alternative, multivariate Bernoulli model, a document is represented by both the absence and presence of features. On a text classification task, McCallum and Nigam (1998) show that the multivariate Bernoulli model performs well with small vocabularies, as opposed to the multinominal model which performs better when larger vocabularies are involved. Larkey (1998) uses the multivariate Bernoulli approach for an essay scoring task, and her results are consistent with the results of McCallum and Nigam (1998) (see also Larkey and Croft (1996) for descriptions of additional applications). In Larkey (1998), sets of essays used for training scoring models typically contain fewer than 300 documents. Furthermore, the vocabulary used across these documents tends to be restricted.

Based on the success of Larkey's experiments, and McCallum and Nigam's findings that the multivariate Bernoulli model performs better on texts with small vocabularies, this approach would seem to be the likely choice when dealing with data sets of essay responses. Therefore, we have adopted this approach in order to build a *thesis statement classifier* that can select from an essay the sentence that is the most likely candidate to be labeled as thesis statement.[2]

In our experiments, we used three general feature types to build the classifier: sentence position; words commonly occurring in thesis statements; and RST labels from outputs generated by an existing rhetorical structure parser (Marcu, 2000).

We trained the classifier to predict thesis statements in an essay. Using the multivariate Bernoulli formula, below, this gives us the log probability that a sentence (S) in an essay belongs to the class (T) of sentences that are thesis statements. We found that it helped performance to use a Laplace estimator to deal with cases where the probability estimates were equal to zero.

$$\log(P(T \mid S)) = \log(P(T)) + \sum_i \begin{cases} \log(P(A_i \mid T)/P(A_i)), & \text{if } S \text{ contains } A_i \\ \log(P(\overline{A}_i \mid T)/P(\overline{A}_i)), & \text{if } S \text{ does not contain } A_i \end{cases}$$

In this formula, P(T) is the prior probability that a sentence is in class T, $P(A_i|T)$ is the conditional probability of a sentence having feature A_i, given that the sentence is in T, and $P(A_i)$ is the prior probability that a sentence contains feature A_i, $P(\overline{A}_i|T)$ is the conditional probability that a sentence does not have feature A_i, given that it is in T, and $P(\overline{A}_i)$ is the prior probability that a sentence does not contain feature A_i.

3.2 Features Used to Classify Thesis Statements

3.2.1 Positional Feature

We found that the likelihood of a thesis statement occurring at the beginning of essays was quite high in the human annotated data. To account for this, we used one feature that reflected the position of each sentence in an essay.

[2] In our research, we trained classifiers using a classical Bayes approach too, where two classifiers were built: a thesis classifier and a non-thesis classifier. In the classical Bayes implementation, each classifier was trained only on positive feature evidence, in contrast to the multivariate Bernoulli approach that trains classifiers both on the absence and presence of features. Since the performance of the classical Bayes classifiers was lower than the performance of the Bernoulli classifier, we report here only the performance of the latter.

3.2.2 Lexical Features

All words from human annotated thesis statements were used to build the Bayesian classifier. We will refer to these words as the *thesis word list*. From the training data, a vocabulary list was created that included one occurrence of each word used in all resolved human annotations of thesis statements. All words in this list were used as independent lexical features. We found that the use of various lists of stop words decreased the performance of our classifier, so we did not use them.

3.2.3 Rhetorical Structure Theory Features

According to RST (Mann and Thompson, 1988), one can associate a rhetorical structure tree to any text. The leaves of the tree correspond to elementary discourse units and the internal nodes correspond to contiguous text spans. Each node in a tree is characterized by a *status* (nucleus or satellite) and *a rhetorical relation*, which is a relation that holds between two non-overlapping text spans. The distinction between nuclei and satellites comes from the empirical observation that the nucleus expresses what is more essential to the writer's intention than the satellite; and that the nucleus of a rhetorical relation is comprehensible independent of the satellite, but not vice versa. When spans are equally important, the relation is multinuclear. Rhetorical relations reflect semantic, intentional, and textual relations that hold between text spans as is illustrated in Figure 2. For example, one text span may elaborate on another text span; the information in two text spans may be in contrast; and the information in one text span may provide background for the information presented in another text span. Figure 2 displays in the style of Mann and Thompson (1988) the rhetorical structure tree of a text fragment. In Figure 2, nuclei are represented using straight lines; satellites using arcs. Internal nodes are labeled with rhetorical relation names.

We built RST trees automatically for each essay using the cue-phrase-based discourse parser of Marcu (2000). We then associated with each sentence in an essay a feature that reflected the status of its parent node (nucleus or satellite), and another feature that reflected its rhetorical relation. For example, for the last sentence in Figure 2 we associated the status satellite and the relation elaboration because that sentence is the satellite of an elaboration relation. For sentence 2, we associated the status nucleus and the relation elaboration because that sentence is the nucleus of an elaboration relation.

We found that some rhetorical relations occurred more frequently in sentences annotated as thesis statements. Therefore, the conditional probabilities for such relations were higher and provided evidence that certain sentences were thesis statements. The Contrast relation shown in Figure 2, for example, was a rhetorical relation that occurred more often in thesis statements. Arguably, there may be some overlap between words in thesis statements, and rhetorical relations used to build the classifier. The RST relations, however, capture long distance relations between text spans, which are not accounted by the words in our thesis word list.

3.3 Evaluation of the Bayesian classifier

We estimated the performance of our system using a six-fold cross validation procedure. We partitioned the 93 essays that were labeled by both human annotators with a thesis statement into six groups. (The judges agreed that 7 of the 100 essays they annotated had no thesis statement.) We trained six times on 5/6 of the labeled data and evaluated the performance on the other 1/6 of the data.

The evaluation results in Table 2 show the average performance of our classifier with respect to the resolved annotation (Alg. wrt. Resolved), using traditional recall (R), precision (P), and F-value (F) metrics. For purposes of comparison, Table 2 also shows the performance of two baselines: the random baseline classifies the thesis statements

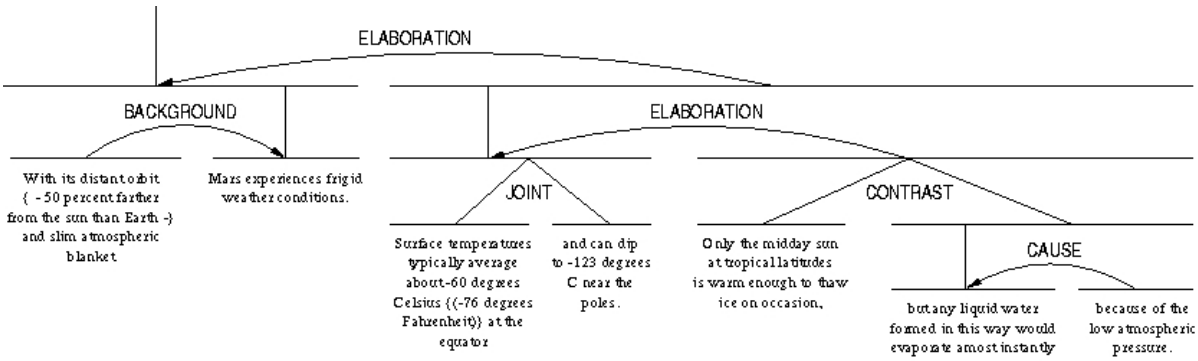

Figure 2: Example of RST tree.

randomly; while the position baseline assumes that the thesis statement is given by the first sentence in each essay.

Table 2: Performance of the thesis statement classifier.

System vs. system	P	R	F
Random baseline wrt. Resolved	0.06	0.05	0.06
Position baseline wrt. Resolved	0.26	0.22	0.24
Alg. wrt. Resolved	0.55	0.46	0.50
1 wrt. 2	0.73	0.69	0.71
1 wrt. Resolved	0.77	0.78	0.78
2 wrt. Resolved	0.68	0.74	0.71

4 Generality of the Thesis Statement Identifier

In commercial settings, it is crucial that a classifier such as the one discussed in Section 3 generalizes across different test questions. New test questions are introduced on a regular basis; so it is important that a classifier that works well for a given data set works well for other data sets as well, without requiring additional annotations and training.

For the thesis statement classifier it was important to determine whether the positional, lexical, and RST-specific features are topic independent, and thus generalizable to new test questions. If so, this would indicate that we could annotate thesis statements across a number of topics, and re-use the algorithm on additional topics, without further annotation. We asked a writing expert to manually annotate the thesis statement in approximately 45 essays for 4 additional test questions: Topics A, C, D and E. The annotator completed this task using the same interface that was used by the two annotators in Experiment 1.

To test generalizability for each of the five EPT questions, the thesis sentences selected by a writing expert were used for building the classifier. Five combinations of 4 prompts were used to build the classifier in each case, and the resulting classifier was then cross-validated on the fifth topic, which was treated as test data. To evaluate the performance of each of the classifiers, agreement was calculated for each 'cross-validation' sample (single topic) by comparing the algorithm selection to our writing expert's thesis statement selections. For example, we trained on Topics A, C, D, and E, using the thesis statements selected manually. This classifier was then used to select, automatically, thesis statements for Topic B. In the evaluation, the algorithm's selection was compared to the manually selected set of thesis statements for Topic B, and agreement was calculated. Table 3 illustrates that in all but one case, agreement exceeds both baselines from Table 2. In this set of manual annotations, the human judge almost always selected one sentence as the thesis statement. This is why Precision, Recall, and the F-value are often equal in Table 3.

Table 3: Cross-topic generalizability of the thesis statement classifier.

Training Topics	CV Topic	P	R	F
ABCD	E	0.36	0.36	0.36
ABCE	D	0.49	0.49	0.49
ABDE	C	0.45	0.45	0.45
ACDE	B	0.60	0.59	0.59
BCDE	A	0.25	0.24	0.25
Mean		0.43	0.43	0.43

5 Discussion and Conclusions

The results of our experimental work indicate that the task of identifying thesis statements in essays is well defined. The empirical evaluation of our algorithm indicates that with a relatively small corpus of manually annotated essay data, one can build a Bayes classifier that identifies thesis statements with good accuracy. The evaluations also provide evidence that this method for automated thesis selection in essays is generalizable. That is, once trained on a few human annotated prompts, it can be applied to other prompts given a similar population of writers, in this case, writers at the college freshman level. The larger implication is that we begin to see that there are underlying discourse elements in essays that can be identified, independent of the topic of the test question. For essay evaluation applications this is critical since new test questions are continuously being introduced into on-line essay evaluation applications.

Our results compare favorably with results reported by Teufel and Moens (1999) who also use Bayes classification techniques to identify rhetorical arguments such as *aim* and *background* in scientific texts, although the texts we are working with are extremely noisy. Because EPT essays are often produced for high-stake exams, under severe time constraints, they are often ungrammatical, repetitive, and poorly organized at the discourse level.

Current investigations indicate that this technique can be used to reliably identify other essay-specific discourse elements, such as, concluding statements, main points of arguments, and supporting ideas. In addition, we are exploring how we can use estimated probabilities as confidence measures of the decisions made by the system. If the confidence level associated with the identification of a thesis statement is low, the system would instruct the student that no explicit thesis statement has been found in the essay.

Acknowledgements

We would like to thank our annotation experts, Marisa Farnum, Hilary Persky, Todd Farley, and Andrea King.

References

Burstein, J., Kukich, K. Wolff, S. Lu, C. Chodorow, M, Braden-Harder, L. and Harris M.D. (1998). Automated Scoring Using A Hybrid Feature Identification Technique. *Proceedings of ACL*, 206-210.

Foltz, P. W., Kintsch, W., and Landauer, T.. (1998). The Measurement of Textual Coherence with Latent Semantic Analysis. *Discourse Processes*, 25(2&3), 285-307.

Grosz B. and Sidner, C. (1986). Attention, Intention, and the Structure of Discourse. Computational Linguistics, 12 (3), 175-204.

Krippendorff K. (1980). *Content Analysis: An Introduction to Its Methodology.* Sage Publ.

Larkey, L. and Croft, W. B. (1996). Combining Classifiers in Text Categorization. *Proceedings of SIGIR*, 289-298.

Larkey, L. (1998). Automatic Essay Grading Using Text Categorization Techniques. *Proceedings of SIGIR*, pages 90-95.

Mani, I. and Maybury, M. (1999). *Advances in Automatic Text Summarization.* The MIT Press.

Mann, W.C. and Thompson, S.A.(1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3), 243–281.

Martin, J. (1992). *English Text. System and Structure.* John Benjamin Publishers.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization.* The MIT Press.

McCallum, A. and Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *The AAAI-98 Workshop on "Learning for Text Categorization".*

Page, E.B. and Peterson, N. (1995). The computer moves into essay grading: updating the ancient test. Phi Delta Kappa, March, 561-565.

Scardamalia, M. and Bereiter, C. (1985). Development of Dialectical Processes in Composition. In Olson, D. R., Torrance, N. and Hildyard, A. (eds), *Literacy, Language, and Learning: The nature of consequences of reading and writing.* Cambridge University Press.

Siegel S. and Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill.

Teufel, S. and Moens, M. (1999). Discourse-level argumentation in scientific articles. Proceedings of the ACL99 Workshop on Standards and Tools for Discourse Tagging.

White E.M. (1994). Teaching and Assessing Writing. Jossey-Bass Publishers, 103-108.