Lexical Chaining for Measuring Discourse Coherence Quality in Test-taker Essays

Swapna Somasundaran
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08540
ssomasundaran@ets.org

Jill Burstein
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08540
jburstein@ets.org

Martin Chodorow
Hunter College, CUNY
695 Park Avenue
New York, NY 10065
martin.chodorow@hunter.cuny.edu

Abstract

This paper presents an investigation of lexical chaining (Morris and Hirst, 1991) for measuring discourse coherence quality in test-taker essays. We hypothesize that attributes of lexical chains, as well as interactions between lexical chains and explicit discourse elements, can be harnessed for representing coherence. Our experiments reveal that performance achieved by our new lexical chain features is better than that of previous discourse features used for this task, and that the best system performance is achieved when combining lexical chaining features with complementary discourse features, such as those provided by a discourse parser based on rhetorical structure theory, and features that reflect errors in grammar, word usage, and mechanics.

1 Introduction

Coherence, the reader's ability to construct meaning from a document, is greatly influenced by the presence and organization of cohesive elements in the text (Halliday and Hasan, 1976; Moe, 1979). The lexical chain (Morris and Hirst, 1991) is one such element. It consists of a sequence of related words that contribute to the continuity of meaning based on word repetition, synonymy and similarity. In this paper we explore how lexical chains can be employed to measure coherence in essays. Specifically, our goal is to investigate how attributes of lexical chains can encode discourse coherence quality, such as adherence to the essay topic, elaboration, usage of varied vocabulary, and sound organization of thoughts and ideas. To do this, we build lexical chains and extract linguistically-motivated features from them. The number of chains and their properties, such as length, density and link strength, can potentially reveal discourse qualities related to focus and elaboration. In addition, features that capture the interactions between chains and explicit discourse cues, such as transition words, can show if the cohesive elements in text have been organized in a coherent fashion.

The main contributions of this paper are as follows: We use lexical chaining features to train a discourse coherence classifier on annotated essays from six different essay-writing tasks which differ in essay genre and/or test-taker population. We then perform experiments to measure the effect of the features when they are used alone and when they are combined with state-of-the-art features to classify the coherence quality of essays. Our results indicate that lexical chaining features yield better results than discourse features previously explored for this task and that the best performing feature combinations contain lexical chaining features. We also show that lexical chaining features can improve system performance across multiple genres of writing and populations. Our efforts result in the creation of a higher performing state-of-the-art feature set for measuring coherence in test-taker writing.

The rest of the paper is organized as follows: In Section 2, we describe our intuitions about lexical chains and how they can be used for measuring discourse coherence quality in essays. Section 3 describes our data, and Section 4 describes our experiments in predicting discourse coherence quality. We discuss related work in Section 5 and conclude in Section 6.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 950–961, Dublin, Ireland, August 23-29 2014.

2 Lexical Chains and Discourse Coherence Quality

According to Morris and Hirst (1991), *lexical cohesion* is the result of chains of related words that contribute to the continuity of lexical meaning. These sequences are characterized by the relations between the words, as well as by their distance and density within a given span. Lexical chains do not stop at sentence boundaries – they can connect a pair of adjacent words or range over an entire text. Morris and Hirst also observe that lexical chains tend to delineate portions of text that have a strong unity of meaning. In this paper, we use this underlying principle of cohesion to detect the *quality of coherence* in a discourse. Specifically, we employ lexical chains to quantify and represent expectations for coherent discourse in test-taker essays. Presumably, violations of these expectations would indicate lack of (or poor) coherence. We believe lexical chains have the potential to reveal the following characteristics about discourse coherence in essays:

Text unity: Textual continuity is vital for the reader's ability to construct meaning from the text (Halliday and Hasan, 1976). Coherent essays generally maintain focus over the main theme, so lexical chains constructed over such essays will have chains representing the central topic running through most of the length of the essay. These types of chains would presumably represent the main claim or position in persuasive texts, the main object or person in descriptive texts, and the main story-line in narrative texts. On the other hand, incoherent texts that jump from one topic to another, or do not adhere to a central idea, will exhibit no chains or chains with very few member words.

Elaboration and Detailing: A function of elaboration in discourse is to overcome misunderstanding or lack of understanding, and to enrich the understanding of the reader by expressing the same thought from a different perspective (Hobbs, 1979). Good writers usually initiate topics, ideas or claims and provide clear elaborations or reasons. That is, a sequence of many related words and phrases will be evoked to explain an idea or provide an account of the writer's reasoning. This development and detailing will be exhibited by lexical chains with a good number of member words.

Variety: While cohesiveness is vital for coherence, too much repetition of the same word can, in fact, harm the discourse quality (Witte and Faigley, 1981). Using a variety of words to express an idea or elaborate on a topic is generally a characteristic of skillful writing. Lexical chains corresponding to such writing will have a variety of similar words within the same chain.

Organization: In addition to cohesion (as represented by lexical chains in our case), one other factor must be present for text to have coherence – organization (Moe, 1979; Perfetti and Lesgold, 1977). Thus, it is important to organize ideas using clear discourse transitions. Transitions from one topic to another, or from a topic to its subtopic, should be clearly cued in order to assist the reader's understanding of the discourse. Consequently, in coherent writing, we would expect lexical chain patterns to synchronize with discourse cues. For example, we would expect some chains to start after a new (sub) topic initiation cue, such as "Secondly" or "Finally", and at least some chains (corresponding to the previous topic) to end immediately before the cue. Similarly, we would expect at least some chains to cross over discourse cues indicating elaboration or reason (e.g. "because") due to topic continuity.

2.1 Features for Measuring Discourse Coherence

In order to measure discourse coherence quality, we create features based on attributes of lexical chains extracted from essays. These features are then used to train a machine learning model, using essays manually labeled for overall discourse coherence quality.

2.1.1 Lexical Chain Construction

Lexical chains in a text are composed of words and terms that are related. Based on Hirst and St-Onge (1995), these relations can be exact repetitions, called *extra-strong relations*, close synonyms, called *strong relations*, or words with weaker semantic relations, called *medium-strong relations*. We implement the lexical chaining program described in Hirst and St. Onge (1995), where if a word or phrase is potentially chainable, it is considered a candidate *node* for existing chains. First, an extra-strong relation is sought throughout all existing chains, and if one is found, the word is added to it. If not, strong relations are sought, but for these, the search scope is limited to the words in any chain that is

no farther away than the previous six sentences in the text; the search ends as soon as a strong relation is found. Finally, if no relationship has yet been found, medium-strong relations are sought with the search scope limited to words in chains that are no farther away than the previous three sentences. If the node cannot be added to any existing chains, it forms its own single-node chain.

In this work, nouns are the focus of the lexical chains. Nouns, adjective-noun and noun-noun structures are identified as potential chain participants. Lin's thesaurus (Lin, 1998) is used to measure similarity between words and phrases. Candidate pairs receiving similarity scores greater than 0.8 are considered to have an extra-strong relationship (word repetition receives a similarity score of 1), pairs with similarity greater than 0.172 are considered to have a strong relation, and pairs with similarity scores greater than 0.099 are considered to have a medium-strong relation. These thresholds were chosen after qualitative inspection of a separate development data set of essays, and are also based on a previous finding (Burstein et al., 2012) that 0.172 is the mean similarity value across different parts of speech in the Lin thesaurus.

We created two feature sets to capture the intuitions described above. The first set, *LEX-1*, encodes the characteristics of text unity, elaboration and variety, while the second, *LEX-2*, encodes organization.

2.1.2 LEX-1 feature set

In order to capture text unity and detailing, we create features such as: *total number of chains in the essay*, *average chain size*, *number (and percentage) of large chains* (chains having more than four nodes are considered to be large chains[1]). As discussed previously, essays that show ample chaining might do so because they adhere to themes and their development, while the presence of large, dense chains might be an indicator that a topic is being discussed in detail. To represent variety, we employ features such as *number (and percentage) of chains that have a variety of words* (chains containing more than one word/phrase type are considered to have variety), as well as *number (and percentage) of large chains with a variety of words*. To encode the characteristics of cohesive relationships, we look at the nature of the links. Examples of these features are: *number and percentage of each link type*, *number (and percentage) of links of each type in large chains as well as in small chains*. Corresponding to each feature that uses counts (e.g. total number of chains) we also created normalized versions of the numbers to account for the essay length. LEX-1 has a total of 38 features.

2.1.3 LEX-2 feature set

LEX-2 features capture the interactions between discourse transitions, indicated by explicit discourse cues, and lexical chaining patterns. For this, we use a discourse cue tagger described in Burstein et al. (1998) that was specifically developed for tagging discourse cues in the essay genre. Using patterns and syntactic rules, the tagger automatically identifies words and phrases used as discourse cues, and assigns them a discourse tag. Each tag has a primary component, indicating whether an argument (or topic) is being initialized (*arg-init*) or developed (*arg-dev*), and a secondary component indicating the specific type of discourse initialization (e.g. *CLAIM*, *SUMMARY*), or development (e.g. *CLAIM*, *CONTRAST*). Examples of the discourse tags and their cues are: *arg-init:SUMMARY* (e.g. *all in all*, *in conclusion*, *in summary*, *overall*), arg-init:TRANSITION (e.g. *let us*), arg-init:PARALLEL (e.g. *firstly*, *similarly*, *finally*), arg-dev:CONTRAST (e.g. *nonetheless*, *however*, *on the contrary*, *rather than*, *even if*), arg-dev:EVIDENCE (e.g. *because of*, *since*), arg-dev:INFERENCE (e.g. *as a result*, *consequently*, *therefore*), arg-dev:DETAIL (e.g. *as well as*, *in this case*, *in addition*, *such as*), arg-dev:REFORMULATION (e.g. *in other words*, *that is*).

For each discourse cue tagged in the text, we replace the cue with its tag and measure the number of chains that (1) start after it, (2) end before it, and (3) continue over it (chains having nodes before and after the tag). We create such features for the tags in the original form (e.g. *arg-dev:DETAIL*), as well as for the primary component alone (e.g. *arg-dev*) and the secondary component alone (e.g. *DETAIL*). This alleviates the data sparseness that we see with certain tags, and results in a total of 138 tags for the LEX-2 feature set.

[1]This number was chosen after inspecting chains in a separate development data set.

3 Data

We use essays from different essay-writing tasks, representing different genres, writing proficiency and populations. Specifically, our essays consist of the following six subsets:

1. *PE-G-N*: **P**ersuasive/**E**xpository essays written by **g**raduate school applicants who are a mix of **n**ative and non-native speakers. (e.g. "As people rely more and more on technology to solve problems, the ability of humans to think for themselves will surely deteriorate. Discuss the extent to which you agree or disagree ... ") [n= 145 essays]

2. *AC-G-N*: **A**rgumentation **c**ritique essays written by **g**raduate school applicants who are a mix of **n**ative and non-native speakers. ("Examine the stated and/or unstated assumptions of the argument. Be sure to explain how the argument depends on the assumptions and what the implications are if the assumptions prove unwarranted ...") [n= 138 essays]

3. *PE-UG-NN*: **P**ersuasive/**E**xpository essays written by **u**ndergraduate and **g**raduate school applicants, who are **n**on-**n**ative speakers. [n= 146 essays]

4. *CS-UG-NN*: **C**ontrastive **s**ummary essays written by **u**ndergraduate and **g**raduate school applicants who are **n**on-**n**ative speakers. Here, the prompt focuses on a specific type of summarization, where ideas from an audio lecture are to be contrasted with ideas from a written passage. [n= 147 essays]

5. *S-G-N*: **S**ubject matter essays written by **g**raduates in a professional licensure exam who are a mix of **n**ative and non-native speakers. [n= 150 essays]

6. *M-K12-N*: A **M**ix of expository, persuasive, descriptive and narrative essays written by **K**-**12** school students who are a mix of **n**ative and non-native speakers. [n= 150 essays]

Of the total of 876 essays, 40 essays were used for system development, and the rest were used for cross-validation experiments. Each essay in the data set was manually annotated for overall discourse coherence quality by annotators not involved in this research. The discourse coherence score was assigned using a 4-point scale (with score point 4 for excellent discourse coherence). Twenty percent of the essays were double annotated and the rest were annotated by one of the annotators. Inter-annotator agreement over the doubly annotated essays, calculated using quadratic weighted kappa (QWK), was 0.61 (substantial agreement). The data distribution for each score point was: 1% for score 1, 9% for score 2, 27% for score 3, 63% for score 4.

4 Experiments

4.1 Baseline Features

A review by Burstein et al. (2013a) describes the several systems that measure discourse coherence quality across various text genres including test-taker essays. Features used to evaluate the discourse coherence quality systems in this study include those previously discussed in Burstein et al. (described below). In addition to comparing our features with previously explored features, our goal is to see if the state-of-the-art feature set can be extended with the use of lexical chaining features.

Entity-grid transition probabilities (entity). Entity-grid transition probabilities (Barzilay and Lapata, 2008) are intended to address unity, progression and coherence by tracking nouns and pronouns in text. An entity grid is constructed in which all entities (nouns and pronouns) are represented by their syntactic roles in a sentence (i.e., Subject, Object, Other). Entity grid transitions track how the same word appears in a syntactic role across adjacent sentences.

Type/Token Ratios for Entities (type/token). These are modified entity-grid transition probabilities. While the entity grid only captures, for example "Subject-Subject" transitions, type/token ratios capture the proportion of unique words that make such transitions. Higher ratios indicate that more concepts are being introduced in a given syntactic role, and lower ratios indicate fewer concepts.

RST-derived features (RST). Rhetorical relations (Mann and Thompson, 1988) derived from an RST parser (Marcu, 2000) are used to evaluate if and how certain rhetorical relations, combinations of rhetorical relations, or rhetorical relation tree structures contribute to discourse coherence quality. These include: (a) relative frequencies of n-gram rhetorical relations in the context of the RST parse tree structure (*unigrams*, or occurrences of a single relation (e.g., ThemeShift); *bigrams*, (e.g., "ThemeShift -> Elaboration"); and *trigrams*, (e.g., "ThemeShift -> Elaboration -> Circumstance")); (b) relative proportions of leaf-parent relation types in the tree structure; and (c) counts of root node relation types in the trees.

Maximum LSA Value for Distant Sentence Pairs (maxLSA). This feature set is the maximum Latent Semantic Analysis (LSA) similarity score found between pairs of sentences that are separated by at least 5 intervening sentences in the essay. It captures reintroduction of topics later in an essay, consistent with a backward inference strategy (Van den Broek et al., 1993; Van den Broek, 2012). LSA has also been employed to measure semantic relatedness between texts for discourse coherence (Foltz et al., 1998).

Grammatical error features (gramErr). These features address errors in grammar that could interfere with a reader's ability to construct meaning and have been used in previous studies (e.g. (Attali and Burstein, 2006; Burstein et al., 2013b)). Specifically, they are based on more than 30 kinds of errors in grammar, such as subject-verb agreement errors, in word usage, such as missing article errors, and in spelling. We use e-rater®, an essay scoring engine developed by Educational Testing Service (ETS), to detect the grammar errors. Aggregate counts of these errors are used as features for predicting discourse coherence.

Program Features (program). This is a single feature for identifying the data type listed in Section 3. Genre and population play an important role with respect to discourse coherence – essays written by more advanced writers, such as those at the graduate level, are typically more coherent than essays written by populations where English is a second language, or by K-12 school students. Note that the program feature is not linguistically motivated – it does not capture the writing construct or a writing skill. However, it is a strong feature as it can reliably bias the system to change its expectations about the discourse quality based on population and task.

4.2 Principal Components Analysis

To reduce the number of lexical chain features, a Principal Components Analysis (PCA) was calculated on an independent set of 6000 essays randomly sampled from the six task types. For 38 LEX-1 features, a 4-component solution accounted for about 0.70 of the feature variance. An 8-component solution explained about 0.30 of the feature variance for the 138 LEX-2 features. (While the variance was lower for this PCA solution, the components were fairly clean.) The component scores were then computed for the 876 essays in our annotated data set. The 4-component scores were used as LEX-1 features, and the 8-component scores were used as LEX-2 features. PCA was used for lexical chaining features in order to reduce the number of features used to build the models rather than using a much larger number of correlated features. PCA was not applied to features from previous work, as we wanted to reproduce their performance.

4.3 Results

A 10-fold cross-validation was run with an unscaled, gradient boosting regressor[2] tuned using quadratic weighted kappa[3]. Specifically, we used the standard Gradient Boosting Regressor in the scikit-learn toolkit[4] (Pedregosa et al., 2011). The learner was trained to assign 4-point coherence quality scores using different combinations of the feature sets described in sections 2.1 and 4.1. In addition to each of the individual features in Section 4.1, we tested two baseline feature combinations: *Baseline-1*, a system using all discourse-based features from Section 4.1, and *Baseline-2*, a system using all features described in Section 4.1.

[2]We experimented with SVMs and Random Forest learners too, but the best results were obtained with the regressor.

[3]The software for the regressor can be found at https://github.com/EducationalTestingService/skll/

[4]http://scikit-learn.org/stable/modules/generated/sklearn.ensemble

Performance was calculated using Quadratic Weighted Kappa (QWK) (Cohen, 1968), which measures the agreement between the system score and the human-annotated coherence score. QWK corrects for chance agreement, and it penalizes large disagreements more than small disagreements. The formula for QWK is as follows:

$$\kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}o_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}e_{ij}}$$

where k is the total number of categories (4 in our case), o_{ij} is the observed value in cell i, j of the confusion matrix between system predictions and human scores, e_{ij} is the expected value for cell i, j, and w_{ij} is the weight given to the discrepancy between $category_i$ and $category_j$. The expected value e_{ij} is calculated as:

$$e_{ij} = \frac{\sum_{j=1}^{k} o_{ij} \sum_{i=1}^{k} o_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} o_{ij}}$$

For quadratic weighted kappa, w_{ij} is calculated as:

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}$$

where i and j are categories, and k is the total number of categories. We use QWK as it is the standard evaluation metric used in automated essay scoring, and it also helps us to compare our results with previous work.

Table 1 reports the results for our proposed features and for each individual feature set investigated in previous work. Here, feature sets explicitly targeting discourse phenomena are grouped under *Discourse-based Features*. The features grouped under *Non-Discourse Features* also capture coherence quality; however they are based on grammatical errors or data type information. The best performing system in each group is shown in **bold**. We see that the full set of lexical chaining features (LEX-1 + LEX-2) is the best performing discourse-based feature set. It performs better than each of the other discourse-based features used alone, and also better than Baseline-1, which uses a combination of *all* discourse-based features from previous work. Notice that the performance of each discourse-based system is below the performance of both *gramErr* and *program*, indicating that they can play an important role in predicting text coherence.

While grammar (*gramErr*) and data type (*program*) are powerful features, it is also important to incorporate capabilities for detecting and evaluating discourse-specific phenomena to ensure construct relevance, as the grading guidelines for essays specify the need for proper organization of ideas (e.g."sustains a well-focused, well-organized analysis, connecting ideas logically"). Lack of construct relevance has been a major criticism of automated scoring methods (Deane, 2013; Shermis and Burstein, 2013). Additionally, discourse-relevant features will allow for interpretable, useful, explicit feedback to students regarding discourse coherence and its breakdown.

In Table 1 we also see that no individual discourse-based system outperforms Baseline-2, comprising all features from the state-of-the-art (Section 4.1). In fact, the human-system agreement obtained by Baseline-2 surpasses the human-human agreement (QWK of 0.61) reported in Section 3. This phenomenon is not uncommon in essay scoring. For example, Bridgeman et al. (2012) performed detailed analyses and found that across all test populations, human-automated system score correlations surpassed human-human score correlations.

Because the *gramErr* and *program* features contain information that is complementary to discourse-based features, we combined the discourse features, first with gramErr features, and then with gramErr+program features. Table 2 reports the results from these experiments. The best performing system for each column is in **bold**, and all features with QWK higher than Baseline-2 are in *italics*. Here,

Feature set	QWK
Discourse-based features	
LEX-1+ LEX-2	**0.316**
LEX-1	0.176
LEX-2	0.246
entity	0.249
type/token	0.178
RST	0.295
maxLSA	0.171
Baseline-1	0.302
Non-Discourse Features	
gramErr	**0.592**
program	0.387
Baseline-2	0.631

Table 1: Performance of individual feature sets.

Feature set	+gramErr	+gramErr +program
LEX-1+ LEX-2	0.608	*0.646*
LEX-1	0.611	*0.650*
LEX-2	0.577	***0.654***
entity	**0.621**	0.609
type/token	0.600	0.623
RST	0.612	*0.649*
maxLSA	0.592	*0.650*
gramErr+program	*0.644*	

Table 2: Performance (QWK), of individual discourse-based features when *gramErr* is added (column 2) and *gramErr* and *program* are added (column 3)

we see that, when combined with *gramErr+program*, the full set of lexical chaining features (LEX-1+LEX-2), as well as LEX-1 and LEX-2 individually, perform above Baseline-2. Surprisingly, we find that when some individual discourse features from previous work are combined with *gramErr+program*, they achieve better performance than Baseline-2 indicating that using the full combination of discourse features may not result in the best system. In the last row, we see that the combination of gramErr and program features alone (*gramErr+program*) is more competitive that Baseline-2, underscoring their usefulness for detecting coherence quality.

Finally, we performed full ablation studies to see which feature set combination produces the best system for identifying discourse coherence quality. Different combinations of the 8 feature sets resulted in 255 different systems, which we ranked based on their performance. Table 3 lists some of the systems, with their respective ranks and QWK values.

First, we observe that the best performing system contains the full set of lexical chaining features and achieves a QWK of 0.673. In fact, all of the top-5 performing systems contain either *LEX-1* or *LEX-2*. The best performance produced by a system not containing any lexical chaining features ranks eighth (*gramErr+ maxLSA+ program+ RST*). Notice that *gramErr+program* is at rank 31, Baseline-2 is at rank 61, and Baseline-1 is at rank 235. Interestingly, RST features are also seen in all of the top-5 systems, indicating that RST features and lexical chaining features capture complementary information about discourse quality. Surprisingly, maxLSA features, which have the same underlying principle of cohesion in text as lexical chains, are in some of the top-performing feature combinations (at ranks 4 and 5), indicating that, in addition to how ideas and themes are presented throughout the essay, the re-introduction of topics is also important.

We tested the statistical significance of the performance differences between our best system (*gramErr + LEX-2+ LEX-1+ maxLSA+ program+ RST*, at rank 1 in Table 3) and three other systems (Baseline-1, Baseline-2 and *gramErr+program*) by drawing 10,000 bootstrap samples (Berg-Kirkpatrick et al., 2012) from our manually scored essays. For each sample, QWKs were calcuated between the human scores and the predictions of our best system, and between the human scores and each of the other three systems' predictions. For each sample, the differences in QWKs were recorded, and the distributions of differences were used for significance testing. Results show that our best performing system is significantly better than Baseline-1 ($p < 0.001$) and Baseline-2 ($p < 0.01$), and it marginally outperformed the system with *gramErr+program* features ($p < 0.06$).

These results show that lexical chaining information is a reliable indicator of discourse quality, and that it can be combined synergistically with other complementary features to extend the state-of-the-art for measuring discourse coherence quality.

Feature set	QWK	Rank
gramErr + LEX-2+ LEX-1+ maxLSA+ program+ RST	**0.673**	1
gramErr+ LEX-1+ program+ RST	0.661	2
gramErr+ LEX-2+ program+ RST	0.661	3
gramErr+ LEX-2+ maxLSA+ program+ RST	0.660	4
gramErr+ LEX-1+ maxLSA+ program+ RST	0.659	5
gramErr+ maxLSA+ program+ RST	0.656	8
gramErr+ program	0.644	31
Baseline-2: entity+ gramErr+ RST+ maxLSA+ program+ type/token	0.631	61
Baseline-1: entity+ RST+ maxLSA+ type/token	0.302	235

Table 3: Performance (QWK), and ranks of systems using different feature combinations

4.4 Analysis by Data Type

In the previous section we saw that features based on lexical chaining are able to successfully encode and predict the quality of discourse coherence. We now examine if this impact is uniform across all essay genres and populations of writers. Table 4 shows the performance of *gramErr +program* (in column 2), the best performing features and their respective performance (**Best system**, columns 3 and 4), and the best feature set when lexical chaining features are removed, with their respective performance (**Best Minus LEX-1 and LEX-2**, columns 5 and 6). Here we use *gramErr +program* as an additional baseline, as it was found to be more competitive than both Baseline-1 and Baseline-2.

Program	gramErr +prog	Best system		Best Minus LEX-1 and LEX-2	
		Features	QWK	Features	QWK
CS-UG-NN	0.418	gramErr+ maxLSA+ RST	**0.523**	gramErr+ maxLSA+ RST	**0.523**
PE-UG-NN	0.406	gramErr + LEX-2 + maxLSA + RST	**0.468**	gramErr	0.406
PE-G-N	0.614	gramErr + LEX-1 + maxLSA	**0.676**	gramErr + maxLSA	0.650
AC-G-N	0.744	gramErr + LEX-2 + maxLSA	**0.839**	gramErr + maxLSA + type/token	0.766
S-G-N	0.414	entity + gramErr+ LEX-1+ RST+ type/token	**0.532**	gramErr+ RST+ type/token	0.487
M-K12-N	0.635	gramErr + LEX-2 + maxLSA	**0.656**	gramErr + maxLSA + RST + type/token	0.649

Table 4: Performance of feature sets by data type. Best performance is shown in **bold**.

In general, for all data types, addition of discourse features produces improvement over just using a combination of *gramErr* and *program* features. Also, the addition of lexical chaining features produces performance improvement for most data types. Specifically, there is substantial improvement in performance for persuasive writing (PE-UG-NN and PE-G-N), expository subject writing (S-G-N) and writing involving critical argumentation (AC-G-N). M-K12-N, which is composed of a mix of genres and writing proficiency, shows a minor improvement. Interestingly, for contrastive summarization (CS-UG-NN), the best system for predicting discourse coherence does not employ any lexical chaining features. For this type of writing, the best feature set using lexical chaining features achieved a QWK of 0.465, which improves over *gramErr+program* but is lower than the best performing feature set. This is perhaps because the discourse phenomena targeted by our lexical chaining features (topical detailing, variety and organization) are already provided for the writer in the source document and the audio lecture, i.e., the materials that are to be referred to in writing this type of essay. Thus, other features play a more prominent role, such as the RST features that capture local discourse organization which is needed, for example, when drawing a contrast between two sources of conflicting information.

5 Related Work

5.1 Discourse coherence quality

A number of models for measuring the quality of discourse coherence have been based on Centering Theory (Grosz et al., 1995). For example, Barzilay and Lapata (2008) construct entity grids based on syntactic subjects and objects. Their algorithm keeps track of the distribution of entity transitions between adjacent sentences and computes a value for all transition types based on their proportion of occurrence in a text. The algorithm has been evaluated with three tasks using well-formed newspaper corpora: text ordering, summary coherence evaluation, and readability assessment. Along similar lines, Rus and Niraula (2012) find centered paragraphs based on prominent syntactic roles. Similarly, Miltsakaki and Kukich (2000) use manually marked centering information and find that higher numbers of Rough Shifts within paragraphs are indicative of a lack of coherence. Using well-formed texts, Pitler and Nenkova (2008) show that a text coherence detection system yields the best performance when it includes features using the Barzilay and Lapata (2008) entity grids, syntactic features, discourse relations from the Penn Discourse Treebank (Prasad et al., 2008), and vocabulary and length features. Wang, Harrington, and White (2012) combine the approaches from Barzilay and Lapata (2008), and Miltsakaki and Kukich (2000) to detect coherence breakdown points. The biggest difference between our approach and the approaches based on Centering Theory is that we do not use syntactically prominent items or try to establish a center. Instead, multiple concurrent thematic chains can "flow" through the paragraph, and their length, density, and interaction with discourse markers are used to model coherence.

In other related work, Lin et al. (2011) use discourse relations from Discourse Lexicalized Tree Adjoining Grammar (D-LTAG) and compile sub-sequences of discourse role transitions to see how the discourse role of a term varies through the progression of the text. Our work, in contrast, traces how chains or thematic threads are organized with respect to the discourse. Our approach also differs from models that measure local coherence between adjacent sentences (Foltz et al., 1998), in that lexical chains can run though the length of the entire text, and hence the features derived from them are able to capture aggregate thematic properties of the entire text such as number, distribution and elaboration of topics.

Discourse coherence models have been previously employed for the task of information-ordering in well-formed texts (e.g., (Soricut and Marcu, 2006; Elsner et al., 2007; Elsner and Charniak, 2008)). In our tasks, discourse coherence quality is influenced by many factors including, but not limited to, ordering of information, such as text unity, detailing and organization.

Higgins et al. (2004) implemented a genre-dependent system to predict discourse coherence quality in essays. Their approach, however, was reliant on organizational structures particular to expository and persuasive essays, such as thesis statement and conclusion.

5.2 Lexical Chaining and Cohesion

Lexical chaining has been used in a number of applications such as news segmentation (Stokes, 2003), question-answering (Moldovan and Novischi, 2002), summarization (Barzilay and Elhadad, 1997), detection and correction of malapropisms (Hirst and St-Onge, 1995), topic detection (Hatch et al., 2000), topic tracking (Carthy and Sherwood-Smith, 2002), and keyword extraction (Ercan and Cicekli, 2007).

In a closely related study, Feng et al. (2009) use lexical chains to measure readability. Lexical chain features are employed to indicate the number of entities/concepts that a reader must keep in mind while reading a document, and two of their features (number of chains in the document and average length of chains) overlap with our LEX-1 features. Our work also differs from systems using cohesion to measure writing quality (e.g., (Witte and Faigley, 1981; Flor et al., 2013)), in that we focus on predicting the quality of discourse coherence.

6 Conclusion

In this paper, we investigated the use of lexical chaining for measuring discourse coherence quality. Based on intuitions about what makes a text coherent, we extracted two sets of features from lexical chains, one encoding how topical themes and cohesive elements are addressed in the text, and another

encoding how the topical themes interact with explicit discourse organizational cues. We performed detailed experiments which showed that lexical chaining features are useful for predicting discourse coherence quality. Specifically, when compared to other previously explored discourse-based features, we found that our lexical chaining features are best performers when used alone. We then experimented with various feature combinations and showed that top performing systems contain lexical chaining features. Our analyses also indicated that lexical chaining features can improve performance on various genres of writing by different populations of writers. Our future work on measuring discourse coherence quality involves extending chains by using verb information and by exploring finer distinctions within the chains themselves (e.g., topical and sub-topical chains).

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v. 2.0. *Journal of Technology, Learning, and Assessment*, 4:3.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL workshop on intelligent scalable text summarization*, volume 17, pages 10–17.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005. Association for Computational Linguistics.

Brent Bridgeman, Catherine Trapani, and Yigal Attali. 2012. Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, 25(1):27–40.

Jill Burstein, Karen Kukich, Susanne Wolff, Ji Lu, and Martin Chodorow. 1998. Enriching automated essay scoring using discourse marking. In *Workshop on Discourse Relations and Discourse Marking*. ERIC Clearinghouse.

Jill Burstein, Jane Shore, John Sabatini, Brad Moulder, Steven Holtzman, and Ted Pedersen. 2012. The language musesm system: Linguistically focused instructional authoring. Technical report, Educational Testing Services (ETS).

Jill Burstein, Joel Tetreault, and Martin Chodorow. 2013a. Holistic discourse coherence annotation for noisy essay writing. *Dialogue and Discourse*, 4(2):34–52.

Jill Burstein, Joel Tetreault, and Nitin Madnani, 2013b. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, chapter The E-rater Automated Essay Scoring System. Routledge.

Joseph Carthy and Michael Sherwood-Smith. 2002. Lexical chains for topic tracking. In *2002 IEEE International Conference on Systems, Man and Cybernetics*, volume 7. IEEE.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213.

Paul Deane. 2013. On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1):7 – 24. Automated Assessment of Writing.

Micha Elsner and Eugene Charniak. 2008. Coreference-inspired coherence modeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 41–44. Association for Computational Linguistics.

Micha Elsner, Joseph Austerweil, and Eugene Charniak. 2007. A unified local and global model for discourse coherence. In *Proceedings of NAACL/HLT*.

Gonenc Ercan and Ilyas Cicekli. 2007. Using lexical chains for keyword extraction. *Information Processing & Management*, 43(6):1705–1714.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–237. Association for Computational Linguistics.

Michael Flor, Beata Beigman Klebanov, and Kathleen M. Sheehan. 2013. Lexical tightness and text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 29–38, Atlanta, Georgia, June. Association for Computational Linguistics.

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Michael AK Halliday and Ruqaiya Hasan. 1976. *Cohesion in english*. English Language Series. Longman Group Ltd.

Paula Hatch, Nicola Stokes, and Joe Carthy. 2000. Topic detection, a new application for lexical chaining. In *the proceedings of BCS-IRSG*, pages 94–103.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *HLT-NAACL*, pages 185–192.

Graeme Hirst and David St-Onge. 1995. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305:305–332.

Jerry R Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3(1):67–90.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 997–1006. Association for Computational Linguistics.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT Press.

Eleni Miltsakaki and Karen Kukich. 2000. Automated evaluation of coherence in student essays. In *Proceedings of LREC 2000*.

Alden J Moe. 1979. Cohesion, coherence, and the comprehension of text. *Journal of Reading*, 23(1):16–20.

Dan Moldovan and Adrian Novischi. 2002. Lexical chains for question answering. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Charles A Perfetti and Alan M Lesgold. 1977. *Discourse Comprehension and Sources of Individual Differences.* ERIC.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse TreeBank 2.0. In *LREC*.

Vasile Rus and Nobal Niraula. 2012. Automated detection of local coherence in short argumentative essays based on centering theory. In *Computational Linguistics and Intelligent Text Processing*, pages 450–461. Springer.

Mark D Shermis and Jill Burstein. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 803–810. Association for Computational Linguistics.

Nicola Stokes. 2003. Spoken and written news story segmentation using lexical chains. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL 2003 student research workshop-Volume 3*, pages 49–54. Association for Computational Linguistics.

Paul Van den Broek, Charles R Fletcher, and Kirsten Risden. 1993. *Investigations of inferential processes in reading: A theoretical and methodological integration*. Taylor & Francis.

Paul Van den Broek. 2012. Individual and developmental differences in reading comprehension: Assessing cognitive processes and outcomes. *Measuring up: Advances in how we assess reading ability*, page 39.

Y Wang, M Harrington, and P White. 2012. Detecting breakdowns in local coherence in the writing of Chinese English learners. *Journal of Computer Assisted Learning*, 28(4):396–410.

Stephen P Witte and Lester Faigley. 1981. Coherence, cohesion, and writing quality. *College composition and communication*, pages 189–204.