Modeling coherence in ESOL learner texts

Helen Yannakoudakis
Computer Laboratory
University of Cambridge
United Kingdom
Helen.Yannakoudakis@cl.cam.ac.uk

Ted Briscoe
Computer Laboratory
University of Cambridge
United Kingdom
Ted.Briscoe@cl.cam.ac.uk

Abstract

To date, few attempts have been made to develop new methods and validate existing ones for automatic evaluation of discourse coherence in the noisy domain of learner texts. We present the first systematic analysis of several methods for assessing coherence under the framework of automated assessment (AA) of learner free-text responses. We examine the predictive power of different coherence models by measuring the effect on performance when combined with an AA system that achieves competitive results, but does not use discourse coherence features, which are also strong indicators of a learner's level of attainment. Additionally, we identify new techniques that outperform previously developed ones and improve on the best published result for AA on a publically-available dataset of English learner free-text examination scripts.

1 Introduction

Automated assessment (hereafter AA) systems of English learner text assign grades based on textual features which attempt to balance evidence of writing competence against evidence of performance errors. Previous work has mostly treated AA as a supervised text classification or regression task. A number of techniques have been investigated, including cosine similarity of feature vectors (Attali and Burstein, 2006), often combined with dimensionality reduction techniques such as Latent Semantic Analysis (LSA) (Landauer et al., 2003), and generative machine learning models (Rudner and Liang, 2002) as well as discriminative ones (Yannakoudakis et al., 2011). As multiple factors influence the linguistic quality of texts, such systems exploit features that correspond to different properties of texts, such as grammar, style, vocabulary usage, topic similarity, and discourse coherence and cohesion.

Cohesion refers to the use of explicit linguistic cohesive devices (e.g., anaphora, lexical semantic relatedness, discourse markers, etc.) within a text that can signal primarily suprasentential discourse relations between textual units (Halliday and Hasan, 1976). Cohesion is not the only mechanism of discourse coherence, which may also be inferred from meaning without presence of explicit linguistic cues. Coherence can be assessed locally in terms of transitions between adjacent clauses, parentheticals, and other textual units capable of standing in discourse relations, or more globally in terms of the overall topical coherence of text passages.

There is a large body of work that has investigated a number of different coherence models on news texts (e.g., Lin et al. (2011), Elsner and Charniak (2008), and Soricut and Marcu (2006)). Recently, Pitler et al. (2010) presented a detailed survey of current techniques in coherence analysis of extractive summaries. To date, however, few attempts have been made to develop new methods and validate existing ones for automatic evaluation of discourse coherence and cohesion in the noisy domain of learner texts, where spelling and grammatical errors are common.

Coherence quality is typically present in marking criteria for evaluating learner texts, and it is iden-

The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pages 33–43,
Montréal, Canada, June 3-8, 2012.

tified by examiners as a determinant of the overall score. Thus we expect that adding a coherence metric to the feature set of an AA system would better reflect the evaluation performed by examiners and improve performance. The goal of the experiments presented in this paper is to measure the effect a number of (previously-developed and new) coherence models have on performance when combined with an AA system that achieves competitive results, but does not use discourse coherence features.

Our contribution is threefold: 1) we present the first systematic analysis of several methods for assessing discourse coherence in the framework of AA of learner free-text responses, 2) we identify new discourse features that serve as proxies for the level of (in)coherence in texts and outperform previously developed techniques, and 3) we improve the best results reported by Yannakoudakis et al. (2011) on the publically available 'English as a Second or Other Language' (ESOL) corpus of learner texts (to date, this is the only public-domain corpus that contains grades). Finally, we explore the utility of our best model for assessing the incoherent 'outlier' texts used in Yannakoudakis et al. (2011).

2 Experimental Design & Background

We examine the predictive power of a number of different coherence models by measuring the effect on performance when combined with an AA system that achieves state-of-the-art results, but does not use discourse coherence features. Specifically, we describe a number of different experiments improving on the AA system presented in Yannakoudakis et al. (2011); AA is treated as a rank preference supervised learning problem and ranking Support Vector Machines (SVMs) (Joachims, 2002) are used to explicitly model the grade relationships between scripts. This system uses a number of different linguistic features that achieve good performance on the AA task. However, these features only focus on lexical and grammatical properties, as well as errors within individual sentences, ignoring discourse coherence, which is also present in marking criteria for evaluating learner texts, as well as a strong indicator of a writer's understanding of a language.

Also, in Yannakoudakis et al. (2011), experiments are presented that test the validity of the system using a number of automatically-created 'outlier' texts. The results showed that the model is vulnerable to input where individually high-scoring sentences are randomly ordered within a text. Failing to identify such pathological cases makes AA systems vulnerable to subversion by writers who understand something of its workings, thus posing a threat to their validity. For example, an examinee might learn by rote a set of well-formed sentences and reproduce these in an exam in the knowledge that an AA system is not checking for prompt relevance or coherence[1].

3 Dataset & Experimental Setup

We use the First Certificate in English (FCE) ESOL examination scripts[2] (upper-intermediate level assessment) described in detail in Yannakoudakis et al. (2011), extracted from the Cambridge Learner Corpus[3] (CLC). The dataset consists of 1,238 texts between 200 and 400 words produced by 1,238 distinct learners in response to two different prompts. An overall mark has been assigned in the range 1–40.

For all experiments, we use a series of 5-fold cross-validation runs on 1,141 texts from the examination year 2000 to evaluate performance as well as generalization of numerous models. Moreover, we identify the best model on year 2000 and we also test it on 97 texts from the examination year 2001, previously used in Yannakoudakis et al. (2011) to report the best published results. Validating the results on a different examination year tests generalization to some prompts not used in 2000, and also allows us to test correlation between examiners and the AA system. Again, we treat AA as a rank preference learning problem and use SVMs, utilizing the SVM^{light} package (Joachims, 2002), to facilitate comparison with Yannakoudakis et al. (2011).

4 Discourse Coherence

We focus on the development and evaluation of (automated) methods for assessing coherence in learner

[1] Powers et al. (2002) report the results of a related experiment with the AA system e-Rater, in which experts tried to subvert the system by submitting essays they believed would be inaccurately scored.

[2] http://ilexir.co.uk/applications/clc-fce-dataset/

[3] http://www.cup.cam.ac.uk/gb/elt/catalogue/subject/custom/item3646603/

texts under the framework of AA. Most of the methods we investigate require syntactic analysis. As in Yannakoudakis et al. (2011), we analyze all texts using the RASP toolkit (Briscoe et al., 2006)[4].

4.1 'Superficial' Proxies

In this section we introduce diverse classes of 'superficial' cohesive features that serve as proxies for coherence. Surface text properties have been assessed in the framework of automatic summary evaluation (Pitler et al., 2010), and have been shown to significantly correlate with the fluency of machine-translated sentences (Chae and Nenkova, 2009).

4.1.1 Part-of-Speech (POS) Distribution

The AA system described in Yannakoudakis et al. (2011) exploited features based on POS tag sequences, but did not consider the distribution of POS types across grades. In coherent texts, constituent clauses and sentences are related and depend on each other for their interpretation. Anaphors such as pronouns link the current sentence to those where the entities were previously mentioned. Pronouns can be directly related to (lack of) coherence and make intuitive sense as cohesive devices. We compute the number of pronouns in a text and use it as a shallow feature for capturing coherence.

4.1.2 Discourse Connectives

Discourse connectives (such as *but* or *because*) relate propositions expressed by different clauses or sentences. The presence of such items in a text should be indicative of (better) coherence. We thus compute a number of shallow cohesive features as proxies for coherence, based on fixed lists of words belonging to the following categories: (a) Addition (e.g., *additionally*), (b) Comparison (e.g., *likewise*), (c) Contrast (e.g., *whereas*) and (d) Conclusion (e.g., *therefore*), and use the frequencies of these four categories as features.

4.1.3 Word Length

The previous AA system treated script length as a normalizing feature, but otherwise avoided such 'superficial' proxies of text quality. However, many cohesive words are longer than average, especially for the closed-class functional component of English vocabulary. We thus assess the minimum, maximum and average word length as a superficial proxy for coherence.

4.2 Semantic Similarity

We explore the utility of inter-sentential feature types for assessing discourse coherence. Among the features used in Yannakoudakis et al. (2011), none explicitly captures coherence and none models inter-sentential relationships. Incremental Semantic analysis (ISA) (Baroni et al., 2007) is a word-level distributional model that induces a semantic space from input texts. ISA is a fully-incremental variation of Random Indexing (RI) (Sahlgren, 2005), which can efficiently capture second-order effects in common with other dimensionality-reduction methods based on singular value decomposition, but does not rely on stoplists or global statistics for weighting purposes.

Utilizing the S-Space package (Jurgens and Stevens, 2010), we trained an ISA model[5] using a subset of ukWaC (Ferraresi et al., 2008), a large corpus of English containing more than 2 billion tokens. We used the POS tagger lexicon provided with the RASP system to discard documents whose proportion of valid English words to total words is less than 0.4; 78,000 documents were extracted in total and were then preprocessed replacing URLs, email addresses, IP addresses, numbers and emoticons with special markers. To measure local coherence we define the similarity between two sentences s_i and s_{i+1} as the maximum cosine similarity between the history vectors of the words they contain. The overall coherence of a text T is then measured by taking the mean of all sentence-pair scores:

$$\text{coherence}(T) = \frac{\sum_{i=1}^{n-1} \max_{k,j} \text{sim}(s_i^k, s_{i+1}^j)}{n-1} \quad (1)$$

where $\text{sim}(s_i^k, s_{i+1}^j)$ is the cosine similarity between the history vectors of the k^{th} word in s_i and the j^{th} word in s_{i+1}, and n is the total number of sentences[6]. We investigate the efficacy of ISA by adding this coherence score, as well as the maximum

[4] http://ilexir.co.uk/applications/rasp/

[5] The parameters of our ISA model are fairly standard: 1800 dimensions, a context window of 3 words, impact rate $i = 0.0003$ and decay rate $k_m = 50$.

[6] We exclude articles, conjunctions, prepositions and auxiliary verbs from the calculation of sentence similarity.

sim value found over the entire text, to the vectors of features associated with a text. The hypothesis is that the degree of semantic relatedness between adjoining sentences serves as a proxy for local discourse coherence; that is, coherent text units contain semantically-related words.

Higgins et al. (2004) and Higgins and Burstein (2007) use RI to determine the semantic similarity between sentences of same/different discourse segments (e.g., from the essay thesis and conclusion, or between sentences and the essay prompt), and assess the percentage of sentences that are correctly classified as related or unrelated. The main differences from our approach are that we assess the utility of semantic space models for predicting the overall grade for a text, in contrast to binary classification at the sentence-level, and we use ISA rather than RI[7].

4.3 Entity-based Coherence

The entity-based coherence model, proposed by Barzilay and Lapata (2008), is one of the most popular statistical models of inter-sentential coherence, and learns coherence properties similar to those employed by Centering Theory (Grosz et al., 1995). Local coherence is modeled on the basis of sequences of entity mentions that are labeled with their syntactic roles (e.g., subject, object). We construct the entity grids using the Brown Coherence Toolkit[8,9] (Elsner and Charniak, 2011b), and use as features the probabilities of different entity transition types, defined in terms of their role in adjacent sentences[10]. Burstein et al. (2010) show how the entity-grid can be used to discriminate high-coherence from low-coherence learner texts. The main difference with our approach is that we evaluate the entity-grid model in the context of AA text grading, rather than binary classification.

[7]We also used RI in addition to ISA, and found that it did not yield significantly different results. In particular, we trained a RI model with 2,000 dimensions and a context window of 3 on the same ukWaC data. Below we only report results for the fully-incremental ISA model.

[8]https://bitbucket.org/melsner/browncoherence

[9]The tool does not perform full coreference resolution; instead, coreference is approximated by linking entities that share a head noun.

[10]We represent entities with specified roles (Subject, Object, Neither, Absent), use transition probabilities of length 2, 3 and 4, and a salience option of 2.

4.4 Pronoun Coreference Model

Pronominal anaphora is another important aspect of coherence. Charniak and Elsner (2009) present an unsupervised generative model of pronominal anaphora for coherence modeling. In their implementation, they model each pronoun as generated by an antecedent somewhere in the previous two sentences. If a 'good' antecedent is found, the probability of a pronoun will be high; otherwise, the probability will be low. The overall probability of a text is then calculated as the probability of the resulting sequence of pronoun assignments. In our experiments, we use the pre-trained model distributed by Charniak and Elsner (2009) for news text to estimate the probability of a text and include it as a feature. However, this model is trained on high-quality texts, so performance may deteriorate when applied to learner texts. It is not obvious how to train such a model on learner texts and we leave this for future research.

4.5 Discourse-new Model

Elsner and Charniak (2008) apply a discourse-new classifier to model coherence. Their classifier distinguishes NPs whose referents have not been previously mentioned in the discourse from those that have been already introduced, using a number of syntactic and lexical features. To model coherence, they assign each NP in a text a label $L_{np} \in \{new, old\}$[11], and calculate the probability of a text as $\Pi_{np:NPs} P(L_{np}|np)$. Again, we use the pre-trained model distributed by Charniak and Elsner (2009) for news text to find the probability of a text following Elsner and Charniak (2008) and include it as a feature.

4.6 IBM Coherence Model

Soricut and Marcu (2006) adapted the IBM model 1 (Brown et al., 1994) used in machine translation (MT) to model local discourse coherence. The intuition behind the IBM model in MT is that the use of certain words in a source language is likely to trigger the use of certain words in a target language. Instead, they hypothesized that the use of certain words in a sentence tends to trigger the use of certain words in an adjoining sentence. In contrast to

[11]NPs with the same head are considered to be coreferent.

semantic space models such as ISA or RI (discussed above), this method models the intuition that local coherence is signaled by the identification of word co-occurrence patterns across adjacent sentences.

We compute two features introduced by Soricut and Marcu (2006): the *forward likelihood* and the *backward likelihood*. The first refers to the likelihood of observing the words in sentence s_{i+1} conditioned on s_i, and the latter to the likelihood of observing the words in s_i conditioned on s_{i+1}. We extract 3 million adjacent sentences from ukWaC[12], and use the GIZA++ (Och and Ney, 2000) implementation of IBM model 1 to obtain the probabilities of recurring patterns. The forward and backward probabilities are calculated over the entire text, and their values are used as features in our feature vectors[13]. We further extend the above model and incorporate syntactic aspects of text coherence by training on POS tags instead of lexical items. We try to model the intuition that local coherence is signaled by the identification of POS co-occurrence patterns across adjacent sentences, where the use of certain POS tags in a sentence tends to trigger the use of other POS tags in an adjacent sentence. We analyze 3 million adjacent sentences using the RASP POS tagger and train the same IBM model to obtain the probabilities of recurring POS patterns.

4.7 Lemma/POS Cosine Similarity

A simple method of incorporating (syntactic) aspects of text coherence is to use cosine similarity between vectors of lemma and/or POS-tag counts in adjacent sentences. We experiment with both: each sentence is represented by a vector whose dimension depends on the total number of lemmas/POS-types. The sentence vectors are weighted using lemma/POS frequency, and the cosine similarity between adjacent sentences is calculated. The coherence of a text T is then calculated as the average value of cosine similarity over the entire text[14]:

$$\text{coherence}(T) = \frac{\sum_{i=1}^{n-1} \text{sim}(s_i, s_{i+1})}{n-1} \quad (2)$$

[12]We use the same subset of documents as the ones used to train our ISA model in Section 4.2.

[13]Pitler et al. (2010) have also investigated the IBM model to measure text quality in automatically-generated texts.

[14]Pitler et al. (2010) use POS cosine similarity to measure continuity in automatically-generated texts.

4.8 Locally-Weighted Bag-of-Words

The popular bag-of-words (BOW) assumption represents a text as a histogram of word occurrences. While computationally efficient, such a representation is unable to maintain any sequential information. The locally-weighted bag-of-words (LOWBOW) framework, introduced by Lebanon et al. (2007), is a sequentially-sensitive alternative to BOW. In BOW, we represent a text as a histogram over the vocabulary used to generate that text. In LOWBOW, a text is represented by a set of local histograms computed across the whole text, but smoothed by kernels centered on different locations.

More specifically, a smoothed characterization of the local histogram is obtained by integrating a length-normalized document with respect to a non-uniform measure that is concentrated around a particular location $\mu \in [0, 1]$. In accordance with the statistical literature on non-parametric smoothing, we refer to such a measure as a smoothing kernel. The kernel parameters μ and σ specify the local histogram's position in the text (i.e., where it is centered) and its scale (i.e., to what extent it is smoothed over the surrounding region) respectively. In contrast to BOW or n-grams, which keep track of frequently occurring patterns independent of their positions, this representation is able to robustly capture medium and long range sequential trends in a text by keeping track of changes in the histograms from its beginning to end.

Geometrically, LOWBOW uses local smoothing to embed texts as smooth curves in the multinomial simplex. These curves summarize the progression of semantic and/or statistical trends through the text. By varying the amount of smoothing we obtain a family of sequential representations possessing different sequential resolutions or scales. Low resolution representations capture topic trends and shifts while ignoring finer details. High resolution representations capture fine sequential details but make it difficult to grasp the general trends within the text[15].

Since coherence involves both cohesive lexical devices and sequential progression within a text, we believe that LOWBOW can be used to assess the sequential content and the global structure and coher-

[15]For more details regarding LOWBOW and its geometric properties see Lebanon et al. (2007).

ence of texts. We use a publically-available LOWBOW implementation[16] to create local histograms over word unigrams. For the LOWBOW kernel smoothing function (see above), we use the Gaussian probability density function restricted to $[0, 1]$ and re-normalized, and a smoothing σ value of 0.02. Additionally, we consider a total number of 9 local histograms (discourse segments). We further extend the above model and incorporate syntactic aspects of text coherence by using local histograms over POS unigrams. This representation is able to capture sequential trends abstracted into POS tags. We try to model the hypothesis that coherence is signaled by sequential, mostly inter-sentential progression of POS types.

Since each text is represented by a set of local histrograms/vectors, and standard SVM kernels cannot work with such input spaces, we use instead a kernel defined over sets of vectors: the diffusion kernel (Lafferty and Lebanon, 2005) compares local histograms in a one-to-one fashion (i.e., histograms at the same locations are compared to each other), and has proven to be useful for related tasks (Lebanon et al., 2007; Escalante et al., 2011). To the best of our knowledge, LOWBOW representations have not been investigated for coherence evaluation (under the AA framework). So far, they have been applied to discourse segmentation (AMIDA, 2007), text categorization (Lebanon et al., 2007), and authorship attribution (Escalante et al., 2011).

5 Evaluation

We examine the predictive power of each of the coherence models/features described in Section 4 by measuring the effect on performance when combined with an AA system that achieves state-of-the-art results on the FCE dataset, but does not use discourse coherence features. In particular, we use the system described in Yannakoudakis et al. (2011) as our baseline AA system. Discourse coherence is a strong indicator of thorough knowledge of a second language and thus we expect coherence features to further improve performance of AA systems.

We evaluate the grade predictions of our models against the gold standard grades in the dataset using Pearson's product-moment correlation coefficient (r) and Spearman's rank correlation coefficient (ρ) as is standard in AA research (Briscoe et al., 2010). Table 1 gives results obtained by augmenting the baseline model with each of the coherence features described above. In each of these experiments, we perform 5-fold cross-validation[17] using all 1,141 texts from the exam year 2000 (see Section 3).

Most of the resulting models have minimal effect on performance[18]. However, word length, ISA, $\text{LOWBOW}_{\text{lex}}$, and the IBM $\text{model}_{\text{POS}_\text{f}}$ derived models all improve performance, while larger differences are observed in r. The highest performance – 0.675 and 0.678 – is obtained with ISA, while the second best feature is word length. The entity-grid, the pronoun model and the discourse-new model do not improve on the baseline. Although these models have been successfully used as components in state-of-the-art systems for discriminating coherent from incoherent news documents (Elsner and Charniak, 2011b), and the entity-grid model has also been successfully applied to learner text (Burstein et al., 2010), they seem to have minimal impact on performance, while the discourse-new model decreases ρ by~0.01. On the other hand, $\text{LOWBOW}_{\text{lex}}$ and $\text{LOWBOW}_{\text{POS}}$ give an increase in performance, which confirms our hypothesis that local histograms are useful. Also, the former seems to perform slightly better than the latter.

Our adapted version of the IBM model – $\text{IBM model}_{\text{POS}}$ – performs better than its lexicalized version, which does not have an impact on performance, while larger differences are observed in r. Additionally, the increase in performance is larger than the one obtained with the entity-grid, pronoun or discourse-new model. The forward version of $\text{IBM model}_{\text{POS}}$ seems to perform slightly better than the backward one, while the results are comparable to $\text{LOWBOW}_{\text{POS}}$ and outperformed by $\text{LOWBOW}_{\text{lex}}$. The rest of the models do not perform as well; the number of pronouns or discourse connectives gives low results, while lemma and POS cosine similarity between adjacent sentences are also

[16] http://goo.gl/yQ0Q0

[17] We compute mean values of correlation coefficients by first applying the r-to-Z Fisher transformation, and then using the Fisher weighted mean correlation coefficient (Faller, 1981).

[18] Significance tests in averaged correlations are omitted as variable estimates are produced, whose variance is hard to be estimated unbiasedly.

		r	ρ
0	Baseline	0.651	0.670
1	POS distr.	0.653	0.670
2	Disc. connectives	0.648	0.668
3	Word length	**0.667**	**0.676**
4	ISA	**0.675**	**0.678**
5	EGrid	0.650	0.668
6	Pronoun	0.650	0.668
7	Disc-new	0.646	0.662
8	$\text{LOWBOW}_{\text{lex}}$	**0.663**	**0.677**
9	$\text{LOWBOW}_{\text{POS}}$	0.659	0.674
10	$\text{IBM model}_{\text{lex}_\text{f}}$	0.649	0.668
11	$\text{IBM model}_{\text{lex}_\text{b}}$	0.649	0.667
12	$\text{IBM model}_{\text{POS}_\text{f}}$	**0.661**	**0.672**
13	$\text{IBM model}_{\text{POS}_\text{b}}$	0.658	0.669
14	Lemma cosine	0.651	0.667
15	POS cosine	0.650	0.665
16	5+6+7+10+11	0.648	0.665
17	All	0.677	0.671

Table 1: 5-fold cross-validation performance on texts from year 2000 when adding different coherence features on top of the baseline AA system.

among the weakest predictors.

Elsner and Charniak (2011b) have shown that combining the entity-grid with the pronoun, discourse-new and lexicalized IBM models gives state-of-the-art results for discriminating news documents and their random permutations. We also combine these models and assess their performance under the AA framework. Row 16 of Table 1 shows that the combination does not give an improvement over the individual models. Moreover, combining all feature classes together in row 17 does not yield higher results than those obtained with ISA, while ρ is no better than the baseline.

In the following experiments, we evaluate the best model identified on year 2000 on a set of 97 texts from the exam year 2001, previously used in Yannakoudakis et al. (2011) to report results of the final best system. Validating the model on a different exam year also shows us the extent to which it generalizes between years. Table 2 presents the results. The published correlations on this dataset are 0.741 and 0.773 r and ρ respectively. Adding ISA on top of the previous system significantly improves[19] the

	r	ρ
Baseline	0.741	0.773
ISA	**0.749**	**0.790***

Table 2: Performance on the exam scripts drawn from the examination year 2001. * indicates a significant improvement at $\alpha = 0.05$.

published results on the 2001 texts, getting closer to the upper-bound. The upper-bound on this dataset[20] is 0.796 and 0.792 r and ρ respectively, calculated by taking the average correlation between the FCE grades and the ones provided by 4 senior ESOL examiners[21]. Table 3 also presents the average correlation between our extended AA system's predicted grades and the 4 examiners' grades, in addition to the original FCE grades from the dataset. Again, our extended model improves over the baseline.

Finally, we explore the utility of our best model for assessing the publically available 'outlier' texts used in Yannakoudakis et al. (2011). The previous AA system is unable to downgrade appropriately 'outlier' scripts containing individually high-scoring sentences with poor overall coherence, created by randomly ordering a set of highly-marked texts. To test our best system, we train an SVM rank preference model with the ISA-derived coherence feature, which can explicitly capture such sequential trends. A generic model for flagging putative 'outlier' texts – whose predicted score is lower than a predefined threshold – for manual checking might be used as the first stage of a deployed AA system. The ISA model improves r and ρ by 0.320 and 0.463 respectively for predicting a score on this type of 'outlier' texts and their original version (Table 4).

6 Analysis & Discussion

In the previous section, we evaluated various cohesion and coherence features on learner data, and found different patterns of performance compared to those previously reported on news texts (see Section 7 for more details). Although most of the models examined gave a minimal effect on AA performance, ISA, $\text{LOWBOW}_{\text{lex}}$, $\text{IBM model}_{\text{POS}_\text{f}}$ and word length

[19] Calculated using one-tailed tests for the difference between dependent correlations (Williams, 1959; Steiger, 1980).

[20] See Yannakoudakis et al. (2011) for details.

[21] The examiners' scores are also distributed with the FCE dataset.

	r	ρ
Baseline	0.723	0.721
ISA	**0.727**	**0.736**

Table 3: Average correlation between the AA model, the FCE dataset grades, and 4 examiners on the exam scripts from year 2000.

	r	ρ
Baseline	0.08	0.163
ISA	**0.400**	**0.626**

Table 4: Performance of the ISA AA model on outliers.

gave a clear improvement in correlation, with larger differences in r. Our results indicate that coherence metrics further improve the performance of a competitive AA system. More specifically, we found the ISA-derived feature to be the most effective contributor to the prediction of text quality. This suggests that incoherence in FCE texts might be due to topic discontinuities. Also, the improvement obtained by LOWBOW suggests that patterns of sequential progression within a text can be useful: coherent texts appear to use similar token distributions at similar locations across different documents.

The word length feature was successfully used as a proxy for coherence, perhaps because many cohesive words are longer than average. However, such a feature can also capture further aspects of texts, such as lexical complexity, so further investigation is needed to identify the extent to which it measures different properties. On the other hand, the minimal effect of the entity-grid, pronoun and discourse-new model suggests that infelicitous use of pronominal forms or sequences of entities may not be an issue in FCE texts. Preliminary investigation of the scripts showed that learners tend to repeat the same entity names or descriptions rather than use pronouns or shorter descriptions.

A possible explanation for the difference in performance between the lexicalized and POS IBM model is that the latter abstracts away from lexical information and thus avoids misspellings and reduces sparsity. Also, our discourse connective classes do not seem to have a predictive power. This may be because our manually-built word lists do not have sufficient coverage.

7 Previous Work

Comparatively few metrics have been investigated for evaluating coherence in (ESOL) learner texts. Miltsakaki and Kukich (2004) employ e-Rater (Attali and Burstein, 2006), an essay scoring system, and show that Centering Theory's Rough-Shift transitions (Grosz et al., 1995) contribute significantly to the assessment of learner texts. Higgins et al. (2004) and Higgins and Burstein (2007) use RI to determine the semantic similarity between sentences of same/different discourse segments. Their model is based on a number of different semantic similarity scores and assesses the percentage of sentences that are correctly classified as (un)related. Among their results, they found that it is hard to beat the baseline (as 98.1% of the sentences were annotated as 'highly related') and identify sentences which are not related to other ones in the same discourse segment. We demonstrate that the related fully-incremental ISA model can be used to improve AA grading accuracy on the FCE dataset, as opposed to classifying the (non-)relatedness of sentences.

Burstein et al. (2010) show how the entity-grid can be used to discriminate high-coherence from low-coherence learner texts. They augment this model with additional features related to writing quality and word usage, and show a positive effect in performance for automated coherence prediction of student essays of different populations. On the FCE dataset used here, entity-grids do not improve AA grading accuracy. This may be because the texts are shorter or because grading is a more difficult task than binary classification. Application of their augmented entity-grid model to FCE texts would be an interesting avenue for future research.

Foltz et al. (1998) examine local coherence in textbooks and articles using Latent Semantic Analysis (LSA) (Landauer et al., 2003). They assess semantic relatedness using vector-based similarity between adjacent sentences. They argue that LSA may be more appropriate for comparing the relative quality of texts; for determining the overall text coherence it may be difficult to set a criterion for the coherence value since it depends on a variety of different factors, such as the size of the text units to be compared. Nevertheless, our results show that ISA, a similar distributional semantic model with dimen-

sionality reduction, improves FCE grading accuracy.

Barzilay and Lee (2004) implement lexicalized content models that represent global text properties on news articles and narratives using Hidden Markov Models (HMMs). In the HMM, states correspond to distinct topics, and transitions between states represent the probability of moving from one topic to another. This approach has the advantage of capturing the order in which different topics appear in texts; however, the HMMs are highly domain specific and would probably need retraining for each distinct essay prompt.

Soricut and Marcu (2006) use a log-linear model that combines local and global models of coherence and show that it outperforms each of the individual ones on news articles and accident reports. Their global model is based on the document content model proposed by Barzilay and Lee (2004). Their local model of discourse coherence is based on the entity-grid (Barzilay and Lapata, 2008), as well as on the lexicalized IBM model (see Section 4.6 above); we have experimented with both, and showed that they have a minimal effect on grading performance with the FCE dataset.

Elsner and Charniak (2008;2011a) apply a discourse-new classifier and a pronoun coreference system to model coherence (see Section 4) on dialogue and news texts. They found that combining these models with the entity-grid achieves state-of-the-art performance. We found that such a combination, as well as the individual models do not perform well for grading the FCE texts.

Recently, Elsner and Charniak (2011a) proposed a variation of the entity-grid intended to integrate topical information. They use Latent Dirichlet Allocation (Blei et al., 2003) to learn topic-to-word distributions, and model coherence by generalizing the binary history features of the entity-grid and computing a real-valued feature which represents the similarity between an entity and the subject(s) of the previous sentence. Also, Lin et al. (2011) proposed a model that assesses the coherence of a text based on discourse relation transitions. The underlying idea is that coherent texts exhibit measurable preferences for specific intra- and inter-discourse relation ordering. They found their model to be complementary to the entity-grid, as it encodes the notion of preferential ordering of discourse relations, and thus tackles local coherence from a different perspective. Applying the above models to AA on learner texts would also be an interesting avenue for future work.

8 Conclusion

We presented the first systematic analysis of a wide variety of models for assessing discourse coherence on learner data, and evaluated their individual performance as well as their combinations for the AA grading task. We adapted the LOWBOW model for assessing sequential content in texts, and showed evidence supporting our hypothesis that local histograms are useful. We also successfully adapted ISA, an efficient and incremental variant distributional semantic model, to this task. ISA, LOWBOW, the POS IBM model and word length are the best individual features for assessing coherence.

A significant improvement over the AA system presented in Yannakoudakis et al. (2011) and the best published result on the FCE dataset was obtained by augmenting the system with an ISA-based local coherence feature. However, it is quite likely that further experimentation with LOWBOW features, given the large range of possible parameter settings, would yield better results too.

We also explored the robustness of the ISA model of local coherence on 'outlier' texts and achieved much better correlations with the examiner's grades for these texts in the FCE dataset. This should facilitate development of an automated system to detect essays consisting of high-quality but incoherent sequences of sentences.

All our results are specific to ESOL FCE texts and may not generalize to other genres or ESOL attainment levels. Future work should also investigate a wider range of (learner) texts and further coherence models, such as that of Elsner and Charniak (2011a) and Lin et al. (2011).

Acknowledgments

We are grateful to Cambridge ESOL, a division of Cambridge Assessment, for supporting this research. We would like to thank Marek Rei and Øistein Andersen for their valuable comments and suggestions, Yi Mao for giving us access to her code, as well as the anonymous reviewers for their useful feedback.

References

AMIDA. 2007. Augmented multi-party interaction with distance access. Available from www. amidaproject.org/, AMIDA Report.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3):1–30.

Marco Baroni, Alessandro Lenci, and Luca Onnis. 2007. ISA meets Lara: An incremental word space model for cognitively plausible simulations of semantic learning. In *Proceedings of the Association for Computational Linguistics*.

Regina Barzilay and Mirella Lapata. 2008. Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics*, 34(1):1–34.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL*, volume 6.

Ted Briscoe, Ben Medlock, and Øistein Andersen. 2010. Automated assessment of ESOL free text examinations. Technical Report UCAM-CL-TR-790, University of Cambridge, Computer Laboratory, November.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1994. The mathematic of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 681–684.

Jieun Chae and Ani Nenkova. 2009. Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 139–147.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 148–156.

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*, pages 41–44.

Micha Elsner and Eugene Charniak. 2011a. Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1179–1189.

Micha Elsner and Eugene Charniak. 2011b. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129.

Hugo J. Escalante, Thamar Solorio, and Manuel Montes-y Gómez. 2011. Local Histograms of Character N-grams for Authorship Attribution. In *Proceedings of the 49th Annual Meeting on Association for Computational Linguistics*, pages 288–298.

Alan J. Faller. 1981. An Average Correlation Coefficient. *Journal of Applied Meteorology*.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In S. Evert, A. Kilgarriff, and S. Sharoff, editors, *Proceedings of the 4th Web as Corpus Workshop*.

Peter W. Foltz, Walter Kintsch, and Thomas K. Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2):285–308.

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 A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English* . Longman Pub Group.

Derrick Higgins and Jill Burstein. 2007. Sentence similarity measures for essay coherence. In *Proceedings of the 7th International Workshop on Computational Semantics*, pages 1–12.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*, pages 133–142.

David Jurgens and Keith Stevens. 2010. The S-Space package: an open source package for word space models. In *Proceedings of the Association for Computational Linguistics 2010 System Demonstrations*, pages 30–35.

John Lafferty and Guy Lebanon. 2005. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6:129–163.

Thomas K. Landauer, Darrell Laham, and Peter W. Foltz. 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M.D. Shermis and J.C. Burstein, editors, *Automated essay scoring: A cross-disciplinary perspective*, pages 87–112.

Guy Lebanon, Yi Mao, and Joshua Dillon. 2007. The locally weighted bag-of-words framework for document representation. *Journal of Machine Learning Research*, 8(10):2405–2441.

Ziheng Lin, Hwee T. Ng, and Min-Yen Kan. 2011. Automatically Evaluating Text Coherence Using Discourse Relations. In *Proceedings of the 49th Annual Meeting on Association for Computational Linguistics*.

Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(01):25–55.

Franz J. Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 544–554.

Donald E. Powers, Jill C. Burstein, Martin Chodorow, Mary E. Fowles, and Karen Kukich. 2002. Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior*, 18(2):103–134.

Lawrence M. Rudner and Tahung Liang. 2002. Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2):3–21.

Magnus Sahlgren. 2005. An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*, pages 1–9. Citeseer.

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.

James H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251.

Evan J. Williams. 1959. The Comparison of Regression Variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 21(2):396–399.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.