University of Pennsylvania
ScholarlyCommons

Departmental Papers (CIS) Department of Computer & Information Science

7-1-2010

Automatic Evaluation of Linguistic Quality in Multi-Document Summarization

Emily Pitler
University of Pennsylvania

Annie Louis
University of Pennsylvania

Ani Nenkova
Univesity of Pennsylvania, nenkova@cis.upenn.edu

Pitler, E., Louis, A., & Nenkova, A., Automatic Evaluation of Linguistic Quality in Multi-Document Summarization, *48th Annual Meeting of the Association for Computational Linguistics*, July 2010, doi: anthology/P10-1056

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/cis_papers/713
For more information, please contact repository@pobox.upenn.edu.

Automatic Evaluation of Linguistic Quality in Multi-Document Summarization

Emily Pitler, Annie Louis, Ani Nenkova
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
epitler,lannie,nenkova@seas.upenn.edu

Abstract

To date, few attempts have been made to develop and validate methods for automatic evaluation of linguistic quality in text summarization. We present the first systematic assessment of several diverse classes of metrics designed to capture various aspects of well-written text. We train and test linguistic quality models on consecutive years of NIST evaluation data in order to show the generality of results. For grammaticality, the best results come from a set of syntactic features. Focus, coherence and referential clarity are best evaluated by a class of features measuring local coherence on the basis of cosine similarity between sentences, coreference information, and summarization specific features. Our best results are 90% accuracy for pairwise comparisons of competing systems over a test set of several inputs and 70% for ranking summaries of a specific input.

1 Introduction

Efforts for the development of automatic text summarizers have focused almost exclusively on improving content selection capabilities of systems, ignoring the linguistic quality of the system output. Part of the reason for this imbalance is the existence of ROUGE (Lin and Hovy, 2003; Lin, 2004), the system for automatic evaluation of content selection, which allows for frequent evaluation during system development and for reporting results of experiments performed outside of the annual NIST-led evaluations, the Document Understanding Conference (DUC)[1] and the Text Analysis Conference (TAC)[2]. Few metrics, however, have been proposed for evaluating linguistic quality and none have been validated on data from NIST evaluations.

In their pioneering work on automatic evaluation of summary coherence, Lapata and Barzilay (2005) provide a correlation analysis between human coherence assessments and (1) semantic relatedness between adjacent sentences and (2) measures that characterize how mentions of the same entity in different syntactic positions are spread across adjacent sentences. Several of their models exhibit a statistically significant agreement with human ratings and complement each other, yielding an even higher correlation when combined.

Lapata and Barzilay (2005) and Barzilay and Lapata (2008) both show the effectiveness of entity-based coherence in evaluating summaries. However, fewer than five automatic summarizers were used in these studies. Further, both sets of experiments perform evaluations of mixed sets of human-produced and machine-produced summaries, so the results may be influenced by the ease of discriminating between a human and machine written summary. Therefore, we believe it is an open question how well these features predict the quality of automatically generated summaries.

In this work, we focus on linguistic quality evaluation for *automatic systems only*. We analyze how well different types of features can rank good and poor machine-produced summaries. Good performance on this task is the most desired property of evaluation metrics during system development. We begin in Section 2 by reviewing the various aspects of linguistic quality that are relevant for machine-produced summaries and currently used in manual evaluations. In Section 3, we introduce and motivate diverse classes of features to capture vocabulary, sentence fluency, and local coherence properties of summaries. We evaluate the predictive power of these linguistic quality metrics by training and testing models on consecutive years of NIST evaluations (data described

[1] http://duc.nist.gov/
[2] http://www.nist.gov/tac/

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 544–554, Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

in Section 4). We test the performance of different sets of features separately and in combination with each other (Section 5). Results are presented in Section 6, showing the robustness of each class and their abilities to reproduce human rankings of systems and summaries with high accuracy.

2 Aspects of linguistic quality

We focus on the five aspects of linguistic quality that were used to evaluate summaries in DUC: grammaticality, non-redundancy, referential clarity, focus, and structure/coherence.[3] For each of the questions, all summaries were manually rated on a scale from 1 to 5, in which 5 is the best.

The exact definitions that were provided to the human assessors are reproduced below.

Grammaticality: The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.

Non-redundancy: There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., "Bill Clinton") when a pronoun ("he") would suffice.

Referential clarity: It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.

Focus: The summary should have a focus; sentences should only contain information that is related to the rest of the summary.

Structure and Coherence: The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

These five questions get at different aspects of what makes a well-written text. We therefore predict each aspect of linguistic quality separately.

3 Indicators of linguistic quality

Multiple factors influence the linguistic quality of text in general, including: word choice, the reference form of entities, and local coherence. We extract features which serve as proxies for each of the factors mentioned above (Sections 3.1 to 3.5). In addition, we investigate some models of grammaticality (Chae and Nenkova, 2009) and coherence (Graesser et al., 2004; Soricut and Marcu, 2006; Barzilay and Lapata, 2008) from prior work (Sections 3.6 to 3.9).

[3] `http://www-nlpir.nist.gov/projects/duc/duc2006/quality-questions.txt`

All of the features we investigate can be computed automatically directly from text, but some require considerable linguistic processing. Several of our features require a syntactic parse. To extract these, all summaries were parsed by the Stanford parser (Klein and Manning, 2003).

3.1 Word choice: language models

Psycholinguistic studies have shown that people read frequent words and phrases more quickly (Haberlandt and Graesser, 1985; Just and Carpenter, 1987), so the words that appear in a text might influence people's perception of its quality. Language models (LM) are a way of computing how familiar a text is to readers using the distribution of words from a large background corpus. Bigram and trigram LMs additionally capture grammaticality of sentences using properties of local transitions between words. For this reason, LMs are widely used in applications such as generation and machine translation to guide the production of sentences. Judging from the effectiveness of LMs in these applications, we expect that they will provide a strong baseline for the evaluation of at least some of the linguistic quality aspects.

We built unigram, bigram, and trigram language models with Good-Turing smoothing over the New York Times (NYT) section of the English Gigaword corpus (over 900 million words). We used the SRI Language Modeling Toolkit (Stolcke, 2002) for this purpose. For each of the three ngram language models, we include the *min*, *max*, and *average* log probability of the sentences contained in a summary, as well as the *overall log probability* of the entire summary.

3.2 Reference form: Named entities

This set of features examines whether named entities have informative descriptions in the summary. We focus on named entities because they appear often in summaries of news documents and are often not known to the reader beforehand. In addition, first mentions of entities in text introduce the entity into the discourse and so must be informative and properly descriptive (Prince, 1981; Fraurud, 1990; Elsner and Charniak, 2008).

We run the Stanford Named Entity Recognizer (Finkel et al., 2005) and record the number of *PERSONs*, *ORGANIZATIONs*, and *LOCATIONs*.

First mentions to people Feature exploration on our development set found that under-specified

references to people are much more disruptive to a summary than short references to organizations or locations. In fact, prior work in Nenkova and McKeown (2003) found that summaries that have been rewritten so that first mentions of people are informative descriptions and subsequent mentions are replaced with more concise reference forms are overwhelmingly preferred to summaries whose entity references have not been rewritten.

In this class, we include features that reflect the modification properties of noun phrases (NPs) in the summary that are first mentions to people. Noun phrases can include pre-modifiers, appositives, prepositional phrases, etc. Rather than pre-specifying all the different ways a person expression can be modified, we hoped to discover the best patterns automatically, by including features for the average number of *each Part of Speech (POS) tag occurring before*, *each syntactic phrase occurring before*[4], *each POS tag occurring after*, and *each syntactic phrase occurring after* the head of the first mention NP for a PERSON. To measure if the lack of pre or post modification is particularly detrimental, we also include the proportion of PERSON first mention NPs *with no words before* and *with no words after* the head of the NP.

Summarization specific Most summarization systems today are *extractive* and create summaries using complete sentences from the source documents. A subsequent mention of an entity in a *source document* which is extracted to be the first mention of the entity in the *summary* is probably not informative enough. For each type of named entity (PERSON, ORGANIZATION, LOCATION), we separately record the number of instances which appear as first mentions in the summary but correspond to non-first mentions in the source documents.

3.3 Reference form: NP syntax

Some summaries might not include people and other named entities at all. To measure how entities are referred to more generally, we include features about the overall syntactic patterns found in NPs: the average number of *each POS tag* and *each syntactic phrase* occurring inside NPs.

[4]We define a linear order based on a preorder traversal of the tree, so syntactic phrases which dominate the head are considered occurring before the head.

3.4 Local coherence: Cohesive devices

In coherent text, constituent clauses and sentences are related and depend on each other for their interpretation. Referring expressions such as pronouns link the current utterance to those where the entities were previously mentioned. In addition, discourse connectives such as “but” or “because” relate propositions or events expressed by different clauses or sentences. Both these categories are known cohesive or linking devices in human-produced text (Halliday and Hasan, 1976). The mere presence of such items in a text would be indicative of better structure and coherence.

We compute a number of shallow features that provide a cheap way of capturing the above intuitions: the number of *demonstratives*, *pronouns*, and *definite descriptions* as well as the number of *sentence-initial discourse connectives*.

3.5 Local coherence: Continuity

This class of linguistic quality indicators is a combination of factors related to coreference, adjacent sentence similarity, and summary-specific context of surface cohesive devices.

Summarization specific Extractive multi-document summaries often lack appropriate antecedents for pronouns and proper context for the use of discourse connectives.

In fact, early work in summarization (Paice, 1980; Paice, 1990) has pointed out that the presence of cohesive devices described in the previous section might in fact be the source of problems. A manual analysis of automatic summaries (Otterbacher et al., 2002) also revealed that anaphoric references that cannot be resolved and unclear discourse relations constitute more than 30% of all revisions required to manually rewrite summaries into a more coherent form.

To identify these potential problems, we adapt the features for surface cohesive devices to indicate whether referring expressions and discourse connectives appear in the summary with the same context as in the input documents.

For each of the cohesive devices discussed in Section 3.4—*demonstratives*, *pronouns*, *definite descriptions*, and *sentence-initial discourse connectives*—we compare the previous sentence in the summary with the previous sentence in the input article. Two features are computed for each type of cohesive device: (1) number of times the preceding sentence in the summary is the same

as the preceding sentence in the input and (2) the number of times the preceding sentence in summary is different from that in the input. Since the previous sentence in the input text often contains the antecedent of pronouns in the current sentence, if the previous sentence from the input is also included in the summary, the pronoun is highly likely to have a proper antecedent.

We also compute the proportion of adjacent sentences in the summary that were extracted from the same input document.

Coreference Steinberger et al. (2007) compare the coreference chains in input documents and in summaries in order to locate potential problems. We instead define a set of more general features related to coreference that are not specific to summarization and are applicable for any text. Our features check the existence of proper antecedents for pronouns in the summary without reference to the text of the input documents.

We use the publicly available pronoun resolution system described in Charniak and Elsner (2009) to mark possible antecedents for pronouns in the summary. We then compute as features the number of times an antecedent for a pronoun was found *in the previous sentence*, *in the same sentence*, or *neither*. In addition, we modified the pronoun resolution system to also output the probability of the most likely antecedent and include the *average antecedent probability* for the pronouns in the text. Automatic coreference systems are trained on human-produced texts and we expect their accuracies to drop when applied to automatically generated summaries. However, the predictions and confidence scores still reflect whether or not possible antecedents exist in previous sentences that match in gender/number, and so may still be useful for coherence evaluation.

Cosine similarity We use cosine similarity to compute the overlap of words in adjacent sentences s_i and s_{i+1} as a measure of continuity.

$$cos\theta = \frac{v_{s_i}.v_{s_{i+1}}}{||v_{s_i}||||v_{s_{i+1}}||} \quad (1)$$

The dimensions of the two vectors (v_{s_i} and $v_{s_{i+1}}$) are the total number of word types from both sentences s_i and s_{i+1}. Stop words were retained. The value of each dimension for a sentence is the number of tokens of that word type in that sentence. We compute the *min*, *max*, and *average* value of cosine similarity over the entire summary.

While some repetition is beneficial for cohesion, too much repetition leads to redundancy in the summary. Cosine similarity is thus indicative of both continuity and redundancy.

3.6 Sentence fluency: Chae and Nenkova (2009)

We test the usefulness of a suite of 38 shallow syntactic features studied by Chae and Nenkova (2009). These features are weakly but significantly correlated with the fluency of machine translated sentences. These include *sentence length*, *number of fragments*, *average lengths of the different types of syntactic phrases*, *total length of modifiers in noun phrases*, and various other syntactic features. We expect that these structural features will be better at detecting ungrammatical sentences than the local language model features.

Since all of these features are calculated over individual sentences, we use the average value over all the sentences in a summary in our experiments.

3.7 Coh-Metrix: Graesser et al. (2004)

The Coh-Metrix tool[5] provides an implementation of 54 features known in the psycholinguistic literature to correlate with the coherence of human-written texts (Graesser et al., 2004). These include commonly used readability metrics based on sentence length and number of syllables in constituent words. Other measures implemented in the system are surface text properties known to contribute to text processing difficulty. Also included are measures of cohesion between adjacent sentences such as similarity under a latent semantic analysis (LSA) model (Deerwester et al., 1990), stem and content word overlap, syntactic similarity between adjacent sentences, and use of discourse connectives. Coh-Metrix has been designed with the goal of capturing properties of coherent text and has been used for grade level assessment, predicting student essay grades, and various other tasks. Given the heterogeneity of features in this class, we expect that they will provide reasonable accuracies for all the linguistic quality measures. In particular, the overlap features might serve as a measure of redundancy and local coherence.

[5] `http://cohmetrix.memphis.edu/`

3.8 Word coherence: Soricut and Marcu (2006)

Word co-occurrence patterns across adjacent sentences provide a way of measuring local coherence that is not linguistically informed but which can be easily computed using large amounts of unannotated text (Lapata, 2003; Soricut and Marcu, 2006). Word coherence can be considered as the analog of language models at the inter-sentence level. Specifically, we used the two features introduced by Soricut and Marcu (2006).

Soricut and Marcu (2006) make an analogy to machine translation: two words are likely to be translations of each other if they often appear in *parallel* sentences; in texts, two words are likely to signal local coherence if they often appear in *adjacent* sentences. The two features we computed are *forward likelihood*, the likelihood of observing the words in sentence s_i conditioned on s_{i-1}, and *backward likelihood*, the likelihood of observing the words in sentence s_i conditioned on sentence s_{i+1}. "Parallel texts" of 5 million adjacent sentences were extracted from the NYT section of GigaWord. We used the GIZA++[6] implementation of IBM Model 1 to align the words in adjacent sentences and obtain all relevant probabilities.

3.9 Entity coherence: Barzilay and Lapata (2008)

Linguistic theories, and Centering theory (Grosz et al., 1995) in particular, have hypothesized that the properties of the transition of attention from entities in one sentence to those in the next, play a major role in the determination of local coherence. Barzilay and Lapata (2008), inspired by Centering, proposed a method to compute the local coherence of texts on the basis of the sequences of entity mentions appearing in them.

In their Entity Grid model, a text is represented by a matrix with rows corresponding to each sentence in a text, and columns to each entity mentioned anywhere in the text. The value of a cell in the grid is the entity's grammatical role in that sentence (Subject, Object, Neither, or Absent). An entity transition is a particular entity's role in two adjacent sentences. The actual entity coherence features are the fraction of each type of these transitions in the entire entity grid for the text. One would expect that coherent texts would contain a certain distribution of entity transitions which would differ from those in incoherent sequences.

We use the Brown Coherence Toolkit[7] (Elsner et al., 2007) to construct the grids. The tool does not perform full coreference resolution. Instead, noun phrases are considered to refer to the same entity if their heads are identical.

Entity coherence features are the only ones that have been previously applied with success for predicting summary coherence. They can therefore be considered to be the state-of-the-art approach for automatic evaluation of linguistic quality.

4 Summarization data

For our experiments, we use data from the multi-document summarization tasks of the Document Understanding Conference (DUC) workshops (Over et al., 2007).

Our training and development data comes from DUC 2006 and our test data from DUC 2007. These were the most recent years in which the summaries were evaluated according to specific linguistic quality questions. Each input consists of a set of 25 related documents on a topic and the target length of summaries is 250 words.

In DUC 2006, there were 50 inputs to be summarized and 35 summarization systems which participated in the evaluation. This included 34 automatic systems submitted by participants, and a baseline system that simply extracted the leading sentences from the most recent article. In DUC 2007, there were 45 inputs and 32 different summarization systems. Apart from the leading sentences baseline, a high performance automatic summarizer from a previous year was also used as a baseline. All these automatic systems are included in our evaluation experiments.

4.1 System performance on linguistic quality

Each summary was evaluated according to the five linguistic quality questions introduced in Section 2: grammaticality, non-redundancy, referential clarity, focus, and structure. For each of these questions, all summaries were manually rated on a scale from 1 to 5, in which 5 is the best.

The distributions of system scores in the 2006 data are shown in Figure 1. Systems are currently the worst at structure, middling at referential clarity, and relatively better at grammaticality, focus,

[6] http://www.fjoch.com/GIZA++.html

[7] http://www.cs.brown.edu/~melsner/manual.html

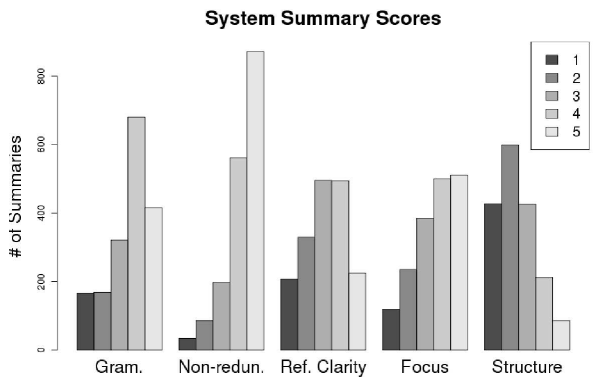

Figure 1: Distribution of system scores on the five linguistic quality questions

	Gram	Non-redun	Ref	Focus	Struct
Content	.02	-.40 *	.29	.28	.09
Gram		.38 *	.25	.24	.54 *
Non-redun			-.07	-.09	.27
Ref				.89 *	.76 *
Focus					.80 *

Table 1: Spearman correlations between the manual ratings for systems averaged over the 50 inputs in 2006; * $p < .05$

and non-redundancy. Structure is the aspect of linguistic quality where there is the most room for improvement. The only system with an average structure score above 3.5 in DUC 2006 was the leading sentences baseline system.

As can be expected, people are unlikely to be able to focus on a single aspect of linguistic quality exclusively while ignoring the rest. Some of the linguistic quality ratings are significantly correlated with each other, particularly referential clarity, focus, and structure (Table 1).

More importantly, the systems that produce summaries with good content[8] are not necessarily the systems producing the most readable summaries. Notice from the first row of Table 1 that none of the system rankings based on these measures of linguistic quality are significantly *positively* correlated with system rankings of content. The development of automatic linguistic quality measurements will allow researchers to optimize both content and linguistic quality.

[8]as measured by summary responsiveness ratings on a 1 to 5 scale, without regard to linguistic quality

5 Experimental setup

We use the summaries from DUC 2006 for training and feature development and DUC 2007 served as the test set. Validating the results on consecutive years of evaluation is important, as results that hold for the data in one year might not carry over to the next, as happened for example in Conroy and Dang (2008)'s work.

Following Barzilay and Lapata (2008), we report summary ranking accuracy as the fraction of correct pairwise rankings in the test set.

We use a Ranking SVM (SVM^{light} (Joachims, 2002)) to score summaries using our features. The Ranking SVM seeks to minimize the number of discordant pairs (pairs in which the gold standard has x_1 ranked strictly higher than x_2, but the learner ranks x_2 strictly higher than x_1). The output of the ranker is always a real valued score, so a global rank order is always obtained. The default regularization parameter was used.

5.1 Combining predictions

To combine information from the different feature classes, we train a meta ranker using the predictions from each class as features.

First, we use a leave-one out (jackknife) procedure to get the predictions of our features for the entire 2006 data set. To predict rankings of systems on one input, we train all the individual rankers, one for each of the classes of features introduced above, on data from the remaining inputs. We then apply these rankers to the summaries produced for the held-out input. By repeating this process for each input in turn, we obtain the predicted scores for each summary.

Once this is done, we use these predicted scores as features for the meta ranker, which is trained on all 2006 data. To test on a new summary pair in 2007, we first apply each individual ranker to get its predictions, and then apply the meta ranker.

In either case (meta ranker or individual feature class), all training is performed on 2006 data, and all testing is done on 2007 data which guarantees the results generalize well at least from one year of evaluation to the next.

5.2 Evaluation of rankings

We examine the predictive power of our features for each of the five linguistic quality questions in two settings. In *system-level* evaluation, we would like to rank all participating systems according to

their performance on the entire test set. In *input-level* evaluation, we would like to rank all summaries produced for a single given input.

For input-level evaluation, the pairs are formed from summaries of the *same input*. Pairs in which the gold standard ratings are tied are not included. After removing the ties, the test set consists of 13K to 16K pairs for each linguistic quality question. Note that there were 45 inputs and 32 automatic systems in DUC 2007. So, there are a total of $45 \cdot \binom{32}{2} = 22,320$ possible summary pairs.

For system-level evaluation, we treat the real-valued output of the SVM ranker for each summary as the linguistic quality score. The 45 individual scores for summaries produced by a given system are averaged to obtain an overall score for the system. The gold-standard system-level quality rating is equal to the *average human ratings* for the system's summaries over the 45 inputs. At the system level, there are about 500 non-tied pairs in the test set for each question.

For both evaluation settings, a random baseline which ranked the summaries in a random order would have an expected pairwise accuracy of 50%.

6 Results and discussion

6.1 System-level evaluation

System-level accuracies for each class of features are shown in Table 2. All classes of features perform well, with at least a 20% absolute increase in accuracy over the random baseline (50% accuracy). For each of the linguistic quality questions, the corresponding best class of features gives prediction accuracies around 90%. In other words, if these features were used to fully automatically compare systems that participated in the 2007 DUC evaluation, only one out of ten comparisons would have been incorrect. These results set a high standard for future work on automatic system-level evaluation of linguistic quality.

The state-of-the-art entity coherence features perform well but are not the best for any of the five aspects of linguistic quality. As expected, sentence fluency is the best feature class for grammaticality. For all four other questions, the best feature set is Continuity, which is a combination of summarization specific features, coreference features and cosine similarity of adjacent sentences. Continuity features outperform entity coherence by 3 to 4% absolute difference on referential quality, focus, and coherence. Accuracies from the language model features are within 1% of entity coherence for these three aspects of summary quality.

Feature set	Gram.	Redun.	Ref.	Focus	Struct.
Lang. models	87.6	83.0	91.2	85.2	86.3
Named ent.	78.5	83.6	82.1	74.0	69.6
NP syntax	85.0	83.8	87.0	76.6	79.2
Coh. devices	82.1	79.5	82.7	82.3	83.7
Continuity	88.8	**88.5**	**92.9**	**89.2**	**91.4**
Sent. fluency	**91.7**	78.9	87.6	82.3	84.9
Coh-Metrix	87.2	86.0	88.6	83.9	86.3
Word coh.	81.7	76.0	87.8	81.7	79.0
Entity coh.	90.2	88.1	89.6	85.0	87.1
Meta ranker	**92.9**	87.9	91.9	87.8	90.0

Table 2: System-level prediction accuracies (%)

Coh-Metrix, which has been proposed as a comprehensive characterization of text, does not perform as well as the language model and the entity coherence classes, which contain considerably fewer features related to only one aspect of text.

The classes of features specific to named entities and noun phrase syntax are the weakest predictors. It is apparent from the results that continuity, entity coherence, sentence fluency and language models are the most powerful classes of features that should be used in automation of evaluation and against which novel predictors of text quality should be compared.

Combining all feature classes with the meta ranker only yields higher results for grammaticality. For the other aspects of linguistic quality, it is better to use Continuity by itself to rank systems.

One certainly unexpected result is that features designed to capture one aspect of well-written text turn out to perform well for other questions as well. For instance, entity coherence and continuity features predict grammaticality with very high accuracy of around 90%, and are surpassed only by the sentence fluency features. These findings warrant further investigation because we would not expect characteristics of local transitions indicative of text structure to have anything to do with sentence grammaticality or fluency. The results are probably due to the significant correlation between structure and grammaticality (Table 1).

6.2 Input-level evaluation

The results of the input-level ranking experiments are shown in Table 3. Understandably, input-level prediction is more difficult and the results are lower compared to the system-level predictions: even with wrong predictions for some of the summaries by two systems, the overall judgment that

one system is better than the other over the entire test set can still be accurate.

While for system-level predictions the meta ranker was only useful for grammaticality, at the input level it outperforms every individual feature class for each of the five questions, obtaining accuracies around 70%.

These input-level accuracies compare favorably with automatic evaluation metrics for other natural language processing tasks. For example, at the 2008 ACL Workshop on Statistical Machine Translation, all fifteen automatic evaluation metrics, including variants of BLEU scores, achieved between 42% and 56% pairwise accuracy with human judgments at the sentence level (Callison-Burch et al., 2008).

As in system-level prediction, for referential clarity, focus, and structure, the best feature class is Continuity. Sentence fluency again is the best class for identifying grammaticality.

Coh-Metrix features are now best for determining redundancy. Both Coh-Metrix and Continuity (the top two features for redundancy) include overlap measures between adjacent sentences, which serve as a good proxy for redundancy.

Surprisingly, the *relative* performance of the feature classes at input level is not the same as for system-level prediction. For example, the language model features, which are the second best class for the system-level, do not fare as well at the input-level. Word co-occurrence which obtained good accuracies at the system level is the least useful class at the input level with accuracies just above chance in all cases.

6.3 Components of continuity

The class of features capturing sentence-to-sentence continuity in the summary (Section 3.5) are the most effective for predicting referential clarity, focus, and structure at the input level. We now investigate to what extent each of its components–summary-specific features, coreference, and cosine similarity between adjacent sentences–contribute to performance.

Results obtained after excluding each of the components of continuity is shown in Table 4; each line in the table represents Continuity minus a feature subclass. Removing cosine overlap causes the largest drop in prediction accuracy, with results about 10% lower than those for the complete Continuity class. Summary specific features, which compare the context of a sentence in the summary with the context in the original document where it appeared, also contribute substantially to the success of the Continuity class in predicting structure and referential clarity. Accuracies drop by about 7% when these features are excluded. However, the coreference features do not seem to contribute much towards predicting summary linguistic quality. The accuracies of the Continuity class are not affected at all when these coreference features are not included.

Feature set	Gram.	Redun.	Ref.	Focus	Struct.
Lang. models	66.3	57.6	62.2	60.5	62.5
Named ent.	52.9	54.4	60.0	54.1	52.5
NP Syntax	59.0	50.8	59.1	54.5	55.1
Coh. devices	56.8	54.4	55.2	52.7	53.6
Continuity	61.7	62.5	**69.7**	**65.4**	**70.4**
Sent. fluency	**69.4**	52.5	64.4	61.9	62.6
Coh-Metrix	65.5	**67.6**	67.9	63.0	62.4
Word coh.	54.7	55.5	53.3	53.2	53.7
Entity coh.	61.3	62.0	64.3	64.2	63.6
Meta ranker	**71.0**	**68.6**	**73.1**	**67.4**	**70.7**

Table 3: Input-level prediction accuracies (%)

6.4 Impact of summarization methods

In this paper, we have discussed an analysis of the outputs of current research systems. Almost all of these systems still use *extractive* methods. The summarization specific continuity features reward systems that include the necessary preceding context from the original document. These features have high prediction accuracies (Section 6.3) of *linguistic quality*, however note that the supporting context could often contain less important *content*. Therefore, there is a tension between strategies for optimizing linguistic quality and for optimizing content, which warrants the development of abstractive methods.

As the field moves towards more *abstractive* summaries, we expect to see differences in both a) summary linguistic quality and b) the features predictive of linguistic aspects.

As discussed in Section 4.1, systems are currently worst at structure/coherence. However, grammaticality will become more of an issue as systems use sentence compression (Knight and Marcu, 2002), reference rewriting (Nenkova and McKeown, 2003), and other techniques to produce their own sentences.

The number of discourse connectives is currently significantly *negatively* correlated with structure/coherence (Spearman correlation of r =

	Ref.	Focus	Struct.
Continuity	69.7	65.4	70.4
- Sum-specific	63.9	64.2	63.5
- Coref	70.1	65.2	70.6
- Cosine	60.2	56.6	60.7

Table 4: Ablation within the Continuity class; pairwise accuracy for input-level predictions (%)

-.06, p = .008 on DUC 2006 system summaries). This can be explained by the fact that they often lack proper context in an extractive summary. However, an *abstractive* system could plan a discourse structure and insert appropriate connectives (Saggion, 2009). In this case, we would expect the presence of discourse connectives to be a mark of a well-written summary.

6.5 Results on human-written abstracts

Since abstractive summaries would have markedly different properties from extracts, it would be interesting to know how well these sets of features would work for predicting the quality of machine-produced abstracts. However, since current systems are extractive, such a data set is not available.

Therefore we experiment on *human-written* abstracts to get an estimate of the expected performance of our features on abstractive system summaries. In both DUC 2006 and DUC 2007, ten NIST assessors wrote summaries for the various inputs. There are four human-written summaries for each input and these summaries were judged on the same five linguistic quality aspects as the machine-written summaries. We train on the human-written summaries from DUC 2006 and test on the human-written summaries from DUC 2007, using the same set-up as in Section 5.

These results are shown in Table 5. We only report results on the input level, as we are interested in distinguishing between the quality of the summaries, not the NIST assessors' writing skills.

Except for grammaticality, the prediction accuracies of the best feature classes for human abstracts are better than those at input level for machine extracts. This result is promising, as it shows that similar features for evaluating linguistic quality will be valid for abstractive summaries as well.

Note however that the relative performance of the feature sets changes between the machine and human results. While for the machines Continuity feature class is the best predictor of referential clarity, focus, and structure (Table 3), for humans, language models and sentence fluency are best for

Feature set	Gram.	Redun.	Ref.	Focus	Struct.
Lang. models	52.1	60.8	76.5	**71.9**	**78.4**
Named ent.	62.5	66.7	47.1	43.9	59.1
NP Syntax	**64.6**	49.0	43.1	49.1	58.0
Coh. devices	54.2	**68.6**	66.7	49.1	64.8
Continuity	54.2	49.0	62.7	61.4	71.6
Sent. fluency	54.2	64.7	**80.4**	**71.9**	72.7
Coh-Metrix	54.2	52.9	68.6	56.1	69.3
Word coh.	62.5	58.8	62.7	70.2	60.2
Entity coh.	45.8	49.0	54.9	52.6	56.8
Meta ranker	62.5	56.9	**80.4**	50.9	67.0

Table 5: Input-level prediction accuracies for human-written summaries (%)

these three aspects of linguistic quality. A possible explanation for this difference could be that in system-produced extracts, incoherent organization influences human perception of linguistic quality to a great extent and so local coherence features turned out very predictive. But in human summaries, sentences are clearly well-organized and here, continuity features appear less useful. Sentence level fluency seems to be more predictive of the linguistic quality of these summaries.

7 Conclusion

We have presented an analysis of a wide variety of features for the linguistic quality of summaries. Continuity between adjacent sentences was consistently indicative of the quality of machine generated summaries. Sentence fluency was useful for identifying grammaticality. Language model and entity coherence features also performed well and should be considered in future endeavors for automatic linguistic quality evaluation.

The high prediction accuracies for input-level evaluation and the even higher accuracies for system-level evaluation confirm that questions regarding the linguistic quality of summaries can be answered reasonably using existing computational techniques. Automatic evaluation will make testing easier during system development and enable reporting results obtained outside of the cycles of NIST evaluation.

Acknowledgments

This material is based upon work supported under a National Science Foundation Graduate Research Fellowship and NSF CAREER award 0953445. We would like to thank Bonnie Webber for productive discussions.

References

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

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106.

J. Chae and A. Nenkova. 2009. Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. In *Proceedings of EACL*, pages 139–147.

E. Charniak and M. Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of EACL*, pages 148–156.

J.M. Conroy and H.T. Dang. 2008. Mind the gap: dangers of divorcing evaluations of summary content from linguistic quality. In *Proceedings of COLING*, pages 145–152.

S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

M. Elsner and E. Charniak. 2008. Coreference-inspired coherence modeling. In *Proceedings of ACL/HLT: Short Papers*, pages 41–44.

M. Elsner, J. Austerweil, and E. Charniak. 2007. A unified local and global model for discourse coherence. In *Proceedings of NAACL/HLT*.

J.R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*, pages 363–370.

K. Fraurud. 1990. Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, 7(4):395.

A.C. Graesser, D.S. McNamara, M.M. Louwerse, and Z. Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods Instruments and Computers*, 36(2):193–202.

B. Grosz, A. Joshi, and S. Weinstein. 1995. Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

K.F. Haberlandt and A.C. Graesser. 1985. Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General*, 114(3):357–374.

M.A.K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman Group Ltd, London, U.K.

T. Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142.

M.A. Just and P.A. Carpenter. 1987. *The psychology of reading and language comprehension*. Allyn and Bacon Boston, MA.

D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*, pages 423–430.

K. Knight and D. Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

M. Lapata and R. Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *International Joint Conference On Artificial Intelligence*, volume 19, page 1085.

M. Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL*, pages 545–552.

C.Y. Lin and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of NAACL/HLT*, page 78.

C.Y. Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, pages 25–26.

A. Nenkova and K. McKeown. 2003. References to named entities: a corpus study. In *Proceedings of HLT/NAACL 2003 (short paper)*.

J. Otterbacher, D. Radev, and A. Luo. 2002. Revisions that improve cohesion in multi-document summaries: a preliminary study. In *Proceedings of the Workshop on Automatic Summarization, ACL*.

P. Over, H. Dang, and D. Harman. 2007. Duc in context. *Information Processing Management*, 43(6):1506–1520.

C.D. Paice. 1980. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, pages 172–191.

C.D. Paice. 1990. Constructing literature abstracts by computer: Techniques and prospects. *Information Processing Management*, 26(1):171–186.

E.F. Prince. 1981. Toward a taxonomy of given-new information. *Radical pragmatics*, 223:255.

H. Saggion. 2009. A Classification Algorithm for Predicting the Structure of Summaries. *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, page 31.

R. Soricut and D. Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of ACL*.

J. Steinberger, M. Poesio, M.A. Kabadjov, and K. Jeek. 2007. Two uses of anaphora resolution in summarization. *Information Processing Management*, 43(6):1663–1680.

A. Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, volume 3.