2010

Structural Features for Predicting the Linguistic Quality of Text: Applications to Machine Translation, Automatic Summarization and Human-Authored Text

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

Jieun Chae
University of Pennsylvania

Annie Louis
University of Pennsylvania

Emily Pitler
University of Pennsylvania

Recommended Citation

Ani Nenkova, Jieun Chae, Annie Louis, and Emily Pitler, "Structural Features for Predicting the Linguistic Quality of Text: Applications to Machine Translation, Automatic Summarization and Human-Authored Text", . January 2010.

Nenkova, A., Chae, J., Louis, A., & Pitler, E., Structural Features for Predicting the Linguistic Quality of Text: Applications to Machine Translation, Automatic Summarization and Human-Authored Text, *Empirical Methods in Natural Language Generation: Data Oriented Methods and Empirical Evaluation*, 2010, doi: 10.1007/978-3-642-15573-4_12

Structural Features for Predicting the Linguistic Quality of Text: Applications to Machine Translation, Automatic Summarization and Human-Authored Text

Abstract

Sentence structure is considered to be an important component of the overall linguistic quality of text. Yet few empirical studies have sought to characterize how and to what extent structural features determine fluency and linguistic quality. We report the results of experiments on the predictive power of syntactic phrasing statistics and other structural features for these aspects of text. Manual assessments of sentence fluency for machine translation evaluation and text quality for summarization evaluation are used as gold-standard. We find that many structural features related to phrase length are weakly but significantly correlated with fluency and classifiers based on the entire suite of structural features can achieve high accuracy in pairwise comparison of sentence fluency and in distinguishing machine translations from human translations. We also test the hypothesis that the learned models capture general fluency properties applicable to human-authored text. The results from our experiments do not support the hypothesis. At the same time structural features and models based on them prove to be robust for automatic evaluation of the linguistic quality of multi-document summaries.

Disciplines

Computer Sciences

Comments

Nenkova, A., Chae, J., Louis, A., & Pitler, E., Structural Features for Predicting the Linguistic Quality of Text: Applications to Machine Translation, Automatic Summarization and Human-Authored Text, *Empirical Methods in Natural Language Generation: Data Oriented Methods and Empirical Evaluation*, 2010, doi: 10.1007/978-3-642-15573-4_12

Structural Features for Predicting the Linguistic Quality of Text

Applications to Machine Translation, Automatic Summarization and Human-Authored Text

Ani Nenkova, Jieun Chae, Annie Louis, and Emily Pitler

University of Pennsylvania
{nenkova,chaeji,lannie,epitler}@seas.upenn.edu

Abstract. Sentence structure is considered to be an important component of the overall linguistic quality of text. Yet few empirical studies have sought to characterize how and to what extent structural features determine fluency and linguistic quality. We report the results of experiments on the predictive power of syntactic phrasing statistics and other structural features for these aspects of text. Manual assessments of sentence fluency for machine translation evaluation and text quality for summarization evaluation are used as gold-standard. We find that many structural features related to phrase length are weakly but significantly correlated with fluency and classifiers based on the entire suite of structural features can achieve high accuracy in pairwise comparison of sentence fluency and in distinguishing machine translations from human translations. We also test the hypothesis that the learned models capture general fluency properties applicable to human-authored text. The results from our experiments do not support the hypothesis. At the same time structural features and models based on them prove to be robust for automatic evaluation of the linguistic quality of multi-document summaries.

1 Introduction

Numerous natural language applications involve the task of producing fluent text. This is a core problem for surface realization in natural language generation [29,2], as well as an important step in machine translation (MT). Considerations of sentence fluency are also key in sentence simplification [42], sentence compression [24,28,11,34,46,18], text re-generation for summarization [6,48] and headline generation [4,49,43]. Despite the popularity of these applications, the factors contributing to sentence fluency have not been researched in depth. Much more attention has been devoted to discourse-level constraints on adjacent sentences indicative of coherence and good text flow [30,5,27]. But the development of fully automatic measures of fluency will make it possible to evaluate system output without the involvement of human assessors, which in turn will facilitate system development.

E. Krahmer, M. Theune (Eds.): Empirical Methods in NLG, LNAI 5790, pp. 222–241, 2010.

In many applications fluency is assessed in combination with other qualities and the assessment is performed in comparison with a human model. For example, in machine translation evaluation, automatic evaluation methods such as BLEU [37] use n-gram overlap comparisons with a model to judge the overall translation quality, with higher n-grams meant to capture fluency considerations. More sophisticated ways to compare a system production and a model involve the use of syntax, but even in these cases fluency is only indirectly assessed and the main advantage of the use of syntax is better estimation of the *semantic* overlap between a model and an output. Similarly, the metrics proposed for text generation by [3] (simple accuracy, generation accuracy) are based on string-edit distance from an ideal output.

In contrast, the work of [48] and [35] directly sets as a goal the assessment of sentence-level fluency, regardless of content and without any human gold-standard. In [48] the main premise is that syntactic information from a parser can more robustly capture fluency than language models, giving more direct indications of the degree of ungrammaticality of a sentence. The idea is extended in [35], where features derived from four different parsers are shown to lead to impressive success in the assessment of fluency of artificially generated sentences with varying level of fluency. Their fluency models hold promise for actual improvements in machine translation output quality [50].

Syntactic tree features that capture common parse configurations and that are used in discriminative parsing [12,9,23] are expected to be beneficial for predicting sentence fluency as well. Indeed, early work has demonstrated that syntactic features, and branching properties in particular, are helpful features for automatically distinguishing human translations from machine translations [15]. The exploration of branching properties of human and machine translations was motivated by the observations during failure analysis that MT system output tends to favor right-branching structures over noun compounding. Branching preference mismatches manifest themselves in the English output when translating from languages whose branching properties are radically different from English. Accuracy close to 80% was achieved for distinguishing human translations from machine translations.

Structural features have also been used for ranking different surface realizations corresponding to the same input semantics, for example in the work of [47] and [8]. In these prior studies, a corpus of English and German sentences respectively are parsed into HPSG/LFG structures. Then all possible surface realizations for the structures are generated and a log-linear model ranker is trained to recognize the original sentence which is considered to be the best realization. Structural features lead to better models than n-gram language model features for both languages. In a follow-up work on human assessment of surface realization variability, Cahill and Forst [7] (this volume) present findings that further motivate the need for automatic objective metric for sentence fluency evaluation. In their experiments, they found that subjects agreed with their own ranking of surface realizations only 70% of the time. A suitable automatic model of fluency will not only be cheaper than manual evaluation but will also remove noise due to human judgement variability.

In our work we continue the investigation of sentence level fluency based on features that capture surface statistics of the syntactic structure in a sentence. We define the features in Sect. 2.1. We revisit the task of distinguishing machine translations from human translations (Sect. 2.3) , but also further our understanding of fluency by providing a comprehensive analysis of the association between fluency assessments of translations and structural features (Sect. 2.2 and Sect. 2.5). We also demonstrate that based on the same class of features, it is possible to distinguish fluent machine translations from non-fluent machine translations (Sect. 2.4). Finally, we test the models on human written text in order to verify if the classifiers trained on data coming from machine translation evaluations can be used for general predictions of fluency and readability (Sect. 3.1 and Sect. 3.2). The results indicate that the models do not generalize well for the different type of data.

Given the findings that fluency models trained on machine translation data do not perform well on human-authored text, we conducted a study where training in testing is performed over the same domain. Specifically, we test the feasibility of performing automatic evaluation of linguistic quality of multi-document summaries using the same structural features (Sect. 4). To ensure that findings are not specific to a given dataset, we train and test the model on consecutive years of evaluations of summarization systems.

2 Sentence Fluency and Machine Translation

For our experiments we use the evaluations of Chinese to English translations distributed by the Linguistic Data Consortium (catalog number LDC2003T17), for which both machine and human translations are available. Machine translations have been assessed by evaluators for fluency on a five point scale (5: flawless English; 4: good English; 3: non-native English; 2: disfluent English; 1: incomprehensible). Assessments by different annotators were averaged to assign overall fluency assessment for each machine-translated sentence. For each segment (sentence), there are four human and three machine translations.

In this setting we address four tasks with increasing difficulty:

- Distinguish human and machine translations.
- Distinguish fluent machine translations from poor machine translations.
- Distinguish the better (in terms of fluency) translation among two translations of the same input segment. This task corresponds to input-level automatic evaluation of fluency.[1]
- Use the models trained on data from MT evaluations to predict potential fluency problems of human-written texts from the Wall Street Journal.

[1] Our data is not suitable for experiments with system-level evaluation where the task is to predict which system is better than others over an entire test suite because there are only three systems. We will address this task for multi-document summarization, where we have summaries produced by 30 or more participating systems.

It is important to note that the purpose of our study is not evaluation of machine translation per se. Our goal is more general and the interest is in finding predictors of sentence fluency. There are no corpora with fluency assessments collected for human-authored text, so it seems advantageous to use the assessments done in the context of machine translation for preliminary investigations of fluency. Nevertheless, our findings are also potentially beneficial for sentence-level evaluation of machine translation.

2.1 Features

Perceived sentence fluency is influenced by many factors. The way the sentence fits in the context of surrounding sentences is one obvious factor [5]. Another well-known factor is vocabulary use: the presence of uncommon difficult words is known to pose problems to readers and to render text less readable [13,41]. But these discourse- and vocabulary-level features measure properties at granularities different from the sentence level.

Structural sentence level features have not been investigated as a stand-alone class, as has been done for the other types of features. This is why we constrain our study to syntactic features alone, and do not initially discuss discourse and language model features in our experiments with machine translation data. For our experiments on evaluation of the linguistic quality of multi-sentential summaries, we do compare several classes of features.

In our work, instead of looking at the syntactic structures present in the sentences, e.g. the syntactic rules used, we use surface statistics of phrase length and types of modification. The sentences were parsed with Charniak's parser [10] in order to calculate these features.

In order to facilitate later reference to features that turn out to be significant in correlation analysis with fluency ratings, we denote some of the Feature Classes by $\textsc{fc}_n$.

Sentence length is the number of words in a sentence. Evaluation metrics such as BLEU [37] have a built-in preference for shorter translations. In general one would expect that shorter sentences are easier to read and thus are perceived as more fluent. We added this feature in order to test directly the hypothesis for brevity preference.

Parse tree depth and the number of subordinating conjunctions (*SBAR count*) are considered to be a measure of sentence complexity, as well as the number of noun phrases, verb phrases and prepositional phrases [38]. Generally, longer sentences are syntactically more complex but when sentences are approximately the same length parse tree depth can be indicative of increased complexity that can slow processing and lead to lower perceived fluency of the sentence.

Number of fragment tags in the sentence parse. Fragments occur without necessarily causing fluency problems in headlines (e.g. "Cheney willing to hold bilateral talks if Arafat observes U.S. cease-fire arrangement") but in machine translation the presence of fragments can signal a more serious problem.

Phrase type proportion was computed for prepositional phrases (PP), noun phrases (NP) and verb phrases (VP). The length in number of words of each phrase type was counted, then divided by the sentence length. Embedded phrases were also included in the calculation: for example a noun phrase (NP1 ... (NP2)) would contribute $length(NP1) + length(NP2)$ to the phrase length count.

Average phrase length is the number of words comprising a given type of phrase, divided by the number of phrases of this type. It was computed for PP, NP, VP, ADJP, ADVP. Two versions of the features were computed— ($\mathbf{FC}_1$) one with embedded phrases included in the calculation and ($\mathbf{FC}_2$) one just for the largest phrases of a given type; the average length of *any* phrase type in a sentence was also calculated. *Normalized average phrase length* ($\mathbf{FC}_3$) is computed for PP, NP and VP and is equal to the average phrase length of given type divided by the sentence length. These were computed only for the largest phrases.

Phrase type rate was also computed for PPs, VPs and NPs and is equal to the number of phrases of the given type that appeared in the sentence, divided by the sentence length. For example, the sentence "The boy caught a huge fish this morning" will have NP phrase number equal to 3/8 and VP phrase number equal to 1/8.

Phrase length. ($\mathbf{FC}_4$) The number of words in a PP, NP, VP, without any normalization; it is computed only for the largest phrases. *Normalized phrase length* is the average phrase length (for VPs, NPs, PPs) divided by the sentence length. This was computed both for ($\mathbf{FC}_5$) longest phrase where embedded phrases of the same type were counted only once and ($\mathbf{FC}_6$) for each phrase regardless of embedding.

Length of NPs/PPs contained in a VP. The average number of words that constitute a NP or PP within a verb phrase, divided by the length of the verb phrase. Similarly, the *length of PP in NP* was computed.

Head noun modifiers. Noun phrases can be very complex, and the head noun can be modified in a variety of ways—pre-modifiers, prepositional phrase modifiers, apposition. The length in words of these modifiers was calculated. Each feature also had a variant in which the modifier length was divided by the sentence length. Finally, two more features on total modification were computed: one was the sum of all modifier lengths, the other the sum of normalized modifier length.

2.2 Feature Analysis

In this section, we analyze the association of the features that we described above and fluency. Note that the purpose of the analysis is not feature selection—all features will be used in the later experiments. Rather, the analysis is performed in order to better understand which factors are predictive of good fluency.

The distribution of fluency scores in the dataset is rather skewed, with the majority of the sentences rated as being of average fluency 3 as can be seen in Table 1.

Table 1. Distribution of fluency scores

Fluency score	Number of sentences
$1 \leq$ fluency < 2	7
$1 \leq$ fluency < 2	295
$2 \leq$ fluency < 3	1789
$3 \leq$ fluency < 4	521
$4 \leq$ fluency < 5	22

Table 2 lists the features for which Pearson's correlation coefficient between the fluency ratings and the values of features was highest.

First of all, fluency and adequacy as given by MT evaluators are highly correlated (0.7). This is surprisingly high, given that separate fluency and adequacy assessments were elicited with the idea that these are qualities of the translations that are independent of each other. Fluency was judged directly by the assessors, while adequacy was meant to assess the content of the sentence compared to a human gold-standard. Yet, the assessments of the two aspects were often the same—readability/fluency of the sentence is important for understanding the sentence. Only after the assessor has understood the sentence can (s)he judge how it compares to the human model. One can conclude then that a model of fluency/readability that will allow systems to produce fluent text is key for developing a successful machine translation system.

The next feature most strongly associated with fluency is sentence length. Shorter sentences are easier and perceived as more fluent than longer ones, which is not surprising. Such preference for brevity has been empirically validated in computational linguistics work both for written text [39] and for utterances in dialog [40] (this volume). Note though that the correlation is actually rather weak. It is only one of various fluency factors and has to be accommodated alongside the possibly conflicting requirements shown by the other features. Still, length considerations reappear at sub-sentential (phrasal) levels as well.

Noun phrase length for example has almost the same correlation with fluency as sentence length does. The longer the noun phrases, the less fluent the sentence is. Long noun phrases take longer to interpret and reduce sentence fluency/readability.

Consider the following example:

- *[The dog]* jumped over the fence and fetched the ball.
- *[The big dog in the corner]* fetched the ball.

The long noun phrase is more difficult to read, especially in subject position. Similarly the length of the verb phrases signals potential fluency problems as can be seen from the examples of human translation in our corpus:[2]

[2] Human translations were not rated for fluency and were considered ideal, as if rated 5. Such assumptions might be too strong. As we will see later, summaries written by people were occasionally rated as being of poor quality by assessors different from the original writer.

- Most of the US allies in Europe publicly *[object to invading Iraq]*$_{VP}$.
- But this *[is dealing against some recent remarks of Japanese financial minister, Masajuro Shiokawa]*$_{VP}$.

VP distance (the average number of words separating two verb phrases) is also negatively correlated with sentence fluency. In machine translations there is the obvious problem that they might not include a verb for long stretches of text. But even in human written text, the presence of more verbs can make a difference in fluency [1]. Consider the following two sentences:

- In his state of the Union address, Putin also **talked** about the national development plan for this fiscal year and the domestic and foreign policies.
- Inside the courtyard of the television station, a reception team of 25 people **was formed to attend** to those who **came to make** donations in person.

The next strongest correlation is with unnormalized verb phrase length. In fact in terms of correlations, in turned out that it was best not to normalize the phrase length features at all. The normalized versions were also correlated with fluency, but the association was lower than for the direct count without normalization.

Parse tree depth is the final feature correlated with fluency with correlation above 0.1.

Table 2. Pearson's correlation coefficient between fluency and different features. P-values are given in parenthesis.

adequacy 0.701 (0.00)	**sentence length** -0.132 (0.00)	**FC$_4$ for NP** -0.124 (0.00)
VP distance -0.116 (0.00)	**FC$_4$ for VP** -0.109 (0.00)	**max tree depth** -0.106 (0.00)
FC$_2$ any phrase -0.105 (0.00)	**FC$_1$ for NP** -0.097 (0.00)	**FC$_1$ for VP** -0.094 (0.00)
SBAR length -0.086 (0.00)	**FC$_2$ for NP** -0.084 (0.00)	**FC$_4$ for PP** -0.082 (0.00)
FC$_1$ for PP -0.070 (0.00)	**SBAR count** -0.069 (0.001)	**PP length in VP** -0.066 (0.001)
FC$_5$ for PP 0.065 (0.001)	**NP length in VP** -0.058 (0.003)	**FC$_6$ PP** -0.054 (0.006)
FC$_6$ for VP 0.054 (0.005)	**PP length in NP** 0.053 (0.006)	**Fragment** -0.049(0.011)

None of the features related to noun modification—apposition length, number of appositions, number of pre-modifiers, etc—were significantly correlated with fluency at the 0.95 confidence level.

2.3 Distinguishing Human from Machine Translations

In this section we use all the features introduced in Section 2.1 for several classification tasks. Note that while we discussed the high correlation between fluency and adequacy, we do not use adequacy in the experiments that we report from here on.

For all experiments we used four of the classifiers in the WEKA machine learning toolkit [22]: decision tree (J48), logistic regression, support vector machines (SMO), and multi-layer perceptron. All results are for 10-fold cross validation.

We extracted the 300 sentences with highest fluency scores, 300 sentences with lowest fluency scores among machine translations and 300 randomly chosen human translations. We then tried the classification task of distinguishing human and machine translations with different fluency quality (highest and lowest fluency score). We expect that low fluency MT will be more easily distinguished from human translation in comparison with machine translations rated as having high fluency. We also ran experiments with the entire dataset, including all human translations and all machine translations regardless of fluency level.

Results are shown in Table 3. Overall the best classifier is the multi-layer perceptron. On the task using all available data of machine and human translations, the classification accuracy is 86.99%. We expected that distinguishing the machine translations from the human ones will be harder when the best translations are used, compared to the worse translations, but this expectation is fulfilled only for the support vector machine classifier.

The high accuracies shown in Table 3 give convincing evidence that the surface structural statistics can distinguish very well between fluent and non-fluent sentences when the examples come from human and machine-produced text respectively. If this is the case, will it be possible to distinguish between good and bad machine translations as well? In order to answer this question, we ran one more binary classification task. The two classes were the 300 machine translations with highest and lowest fluency respectively. The results are not as good as those for distinguishing machine and human translation, but still significantly outperform a random baseline. All classifiers performed similarly on the task, and achieved accuracy close to 61%.

Table 3. Accuracy for the task of distinguishing machine and human translations

Classifier	worst 300 MT	best 300 MT	all MT
SMO	86.00%	78.33%	82.68%
Logistic reg.	77.16%	79.33%	82.68%
MLP	78.00%	82%	86.99%
Decision Tree(J48)	71.67 %	81.33%	86.11%

2.4 Pairwise Fluency Comparisons

We also considered the possibility of pairwise comparisons for fluency: given two sentences, can we distinguish which is the one scored more highly for fluency. The feature vector for each pair of sentences is obtained as the difference of features of the individual sentences.

There are two ways this task can be set up. First, we can use all assessed translations and make pairings for every two sentences with different fluency assessment. In this setting, the question being addressed is *Can sentences with differing fluency be distinguished?*, without regard to the sources of the sentence. The harder question is *Can a more fluent translation be distinguished from a less fluent translation of the same sentence?*

The results from these experiments can be seen in Table 4. When any two sentences with different fluency assessments are paired, the prediction accuracy is very high: 91.34% for the multi-layer perceptron classifier. In fact all classifiers have accuracy higher than 80% for this task. The surface statistics of syntactic form are powerful enough to distinguishing sentences of varying fluency.

The task of pairwise comparison for translations of the same input is more difficult: doing well on this task would be equivalent to having a reliable measure for ranking different possible translation variants.

Table 4. Accuracy for pairwise fluency comparison. "Same sentence" are comparisons constrained between different translations of the same sentences, "Any pair" contains comparisons of sentences with different fluency over the entire dataset.

Task	J48	Logistic Regression	SMO	MLP
Any pair	89.73%	82.35%	82.38%	91.34%
Same Sentence	67.11%	70.91%	71.23%	69.18%

In fact, the problem is *much* more difficult as can be seen in the second row of Table 4, and the performance for all classifiers is more than 10% lower than those for comparisons not constrained to be translations of the same sentence. Logistic regression, support vector machines and multi-layer perceptron perform similarly, with support vector machine giving the best accuracy of 71.23%. This number is still impressively high, and significantly higher than baseline performance.

2.5 Feature Analysis: Differences among Tasks

In the previous sections we presented three variations involving fluency predictions based on syntactic phrasing features: distinguishing human from machine translations, distinguishing good machine translations from bad machine translations, and pairwise ranking of sentences with different fluency. The results differ considerably and it is interesting to know whether the same kind of features are useful in making the three distinctions.

In Table 5 we show the five features with largest weight in the support vector machine model for each task. In many cases, certain features appear to be important only for particular tasks. For example the number of prepositional phrases is an important feature only for ranking different versions of *the same sentence* but is not important for other distinctions. The number of appositions is helpful in distinguishing human translations from machine translations, but is

Table 5. The five features with highest weights in the support vector machine model for the different tasks

MT vs HT	good MT vs Bad MT	Ranking	Same sentence Ranking
FC_4 for PP	# of SBARs	FC_2 for NP	FC_5 for NP
PP length in VP	FC_4 for VP	FC_3 for PP	# of PP
FC_2 for NP	post modification length	# of NP	FC_6 for NP
# of appositions	# of VP	FC_3 for NP	max tree depth
SBAR length	sentence length	FC_3 for VP	FC_2 any

not that useful in the other tasks. So the predictive power of the features is very directly related to the variant of fluency distinctions one is interested in making.

3 Applications to Human-Authored Text

3.1 Identifying Hard-to-Read Sentences in Wall Street Journal Texts

The goal we set out in the beginning of this paper was to derive a predictive model of sentence fluency from data coming from MT evaluations. In the previous sections, we demonstrated that indeed structural features can enable us to perform this task very accurately *in the context of machine translation.* But will the models conveniently trained on data from MT evaluation be at all capable to identify sentences in human-written text that are not fluent and are difficult to understand?

To answer this question, we performed an additional experiment on 30 Wall Street Journal articles from the Penn Treebank that were previously used in experiments for assessing overall text quality [39]. The articles were chosen at random and comprised a total of 290 sentences. One human assessor was asked to read each sentence and mark the ones that seemed disfluent because they were hard to comprehend. These were sentences that needed to be read more than once in order to fully understand the information conveyed in them. There were 52 such sentences. The assessments served as a gold-standard against which the predictions of the fluency models were compared.

Two models trained on machine translation data were used to predict the status of each sentence in the WSJ articles. One of the models was that for distinguishing human translations from machine translations (human vs. MT), the other was the model for distinguishing the 300 best from the 300 worst machine translations (good MT vs. bad MT). The classifiers used were decision trees for human vs. machine distinction and support vector machines for good MT vs. bad MT. For the first model sentences predicted to belong to the "human translation" class are considered fluent; for the second model fluent sentences are the ones predicted to be in the "good MT" class.

The results are shown in Table 6. The two models differ in performance considerably. The model for distinguishing machine translations from human translations is the better one, with accuracy of 57%. For both, prediction accuracy is much lower than when tested on data from MT evaluations. These findings indicate that building a new corpus for the finer fluency distinctions present in human-written text is likely to be more beneficial than trying to leverage data from existing MT evaluations.

Table 6. Accuracy, precision and recall (for fluent class) for each model when test on WSJ sentences

Model	Accuracy	Precision	Recall
human vs machine trans.	57%	0.79	0.58
good MT vs bad MT	44%	0.57	0.44

Below, we show several example sentences on which the assessor and the model for distinguishing human and machine translations (dis)agreed.

1. Model and assessor agree that sentence is problematic.
   (a) The Soviet legislature approved a 1990 budget yesterday that halves its huge deficit with cuts in defense spending and capital outlays while striving to improve supplies to frustrated consumers.
   (b) Officials proposed a cut in the defense budget this year to 70.9 billion rubles (US$114.3 billion) from 77.3 billion rubles (US$125 billion) as well as large cuts in outlays for new factories and equipment.
   (c) Rather, the two closely linked exchanges have been drifting apart for some years, with a nearly five-year-old moratorium on new dual listings, separate and different listing requirements, differing trading and settlement guidelines and diverging national-policy aims.
2. The model predicts the sentence is good, but the assessor finds it problematic.
   (a) Moody's Investors Service Inc. said it lowered the ratings of some $145 million of Pinnacle debt because of "accelerating deficiency in liquidity," which it said was evidenced by Pinnacle's elimination of dividend payments.
   (b) Sales were higher in all of the company's business categories, with the biggest growth coming in sales of foodstuffs such as margarine, coffee and frozen food, which rose 6.3%.
   (c) Ajinomoto predicted sales in the current fiscal year ending next March 31 of 480 billion yen, compared with 460.05 billion yen in fiscal 1989.
3. The model predicts the sentences are bad, but the assessor considered them fluent.
   (a) The sense grows that modern public bureaucracies simply don't perform their assigned functions well.
   (b) Amstrad PLC, a British maker of computer hardware and communications equipment, posted a 52% plunge in pretax profit for the latest year.
   (c) At current allocations, that means EPA will be spending $300 billion on itself.

3.2 Correlation with Overall Text Quality

Here we focus on the relationship between sentence fluency and overall text quality. We would expect that the presence of disfluent sentences in text will make it appear less well written. Five annotators had previously assessed the overall text quality of each of the WSJ articles on a scale from 1 to 5 [39]. The average of the assessments was taken as a single number describing the linguistic quality article. The correlation between this number and the percentage of fluent sentences in the article according to the different models is shown in Table 7.

The correlation between the percentage of fluent sentences in the article as given by the human assessor and the overall text quality is rather low, 0.127. Correlation with the percentage of fluent sentences predicted by the two automatic models are even closer to zero. Note that none of the correlations are actually significant for the small dataset of 30 points.

Table 7. Correlations between text quality assessment of the articles and the percentage of fluent sentences according to different models

Fluency given by	Correlation	p-value
human	0.127	0.504
human vs machine trans. model	-0.055	0.772
good MT vs bad MT model	0.076	0.69

The low correlations indicate that binary decisions on sentence level fluency are not likely to be helpful for determining the overall quality of text. A question that remains unanswered from the experiments presented so far is whether structural features can be used to predict overall text quality directly. A dataset larger than the 30 WSJ documents is necessary for this purpose. So, in the next section we turn to a large collection of multi-document summaries evaluated for linguistic quality.

4 Predicting Linguistic Quality for Multi-document Summarization

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 [32,33], the system for automatic evaluation of content selection, which allows for frequent system evaluation during system development and for reporting results of experiments performed outside of the annual NIST-led evaluations (DUC[3] and TAC[4]). Few metrics, however, have been proposed

[3] http://duc.nist.gov/

[4] http://www.nist.gov/tac/

[31] for evaluating linguistic quality and none have been tested for correlation with the manual metrics used by NIST.

So here we use the same structural features described in the experiments on sentence level fluency in order to directly predict the linguistic quality of summaries. We compare their performance with that of several other metrics of text quality. We evaluate the predictive power of the linguistic quality metrics by training and testing models on consecutive years of NIST evaluations, showing the robustness of each class and their abilities to reproduce human rankings of systems and summaries with high accuracy.

4.1 Summarization Data

We use a large corpus of system- and human-authored summaries from the Document Understanding Conference (DUC) workshops [36] from years 2006 and 2007. These summaries were produced for inputs consisting of a set of 25 related documents on a topic. The length of the summary was constrained to be 250 or fewer words. In DUC 2006, there were 50 inputs to be summarized and 35 summarization systems which participated in the evaluation. In DUC 2007, there were 45 inputs and 32 different summarization systems. Four human summaries are also available for each input.

All summaries were manually evaluated for several aspects of linguistic quality, including *(a)* referential clarify, *(b)* focus and *(c)* structure and coherence. For each of the questions, Summaries were rated on a scale from 1 to 5, in which 5 is the best separately for each of these aspects.

Judging from the 2006 scores, systems are currently the worst at structure (mean=2.4, median=2), middling at referential clarity (mean=3.1, median=3), and relatively better at focus (mean=3.6, median=4). Structure is the aspect of linguistic quality where there is the most room for improvement. Excluding the baseline system, which simply extracts the leading sentences from the most recent article in the input and therefore has well-formed summaries, all of the other systems have average structure scores below 3.5 in DUC 2006. Human summaries were predominantly scored 5, but some scores of 4 and 3 also occur.

4.2 Predictors of Linguistic Quality

Structural features. The structural features we described in Sect. 2.1 apply for individual sentences. In order to apply they to summaries which consist of more than one sentence, we simply take the average value of features for the sentences in the summary.

Coh-Metrix. 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 [19]. These include for example commonly used readability metrics based on sentence length and number of syllables in constituent words.

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

Other measures implemented in the system are surface text properties known to contribute to text processing difficulty such as the number of words before the main verb, the prevalence of pronouns and low frequency content words. Also included are measures of cohesion between adjacent sentences such as similarity under a latent semantic analysis model [16], stem and content word overlap, and syntactic similarity between adjacent sentences. In addition, the presence in a text of different types of discourse connectives such as *causal (e.g. 'because', 'consequently')* and *temporal (e.g. 'after', 'until')* are also recorded. 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, identifying differences between spoken and written texts, authorship identification, and various other tasks.

Vocabulary: language models. Psycholinguistic studies have shown that people read frequent words and phrases more quickly [26,21], so the words that appear in a text might influence people's perception of its quality. Language models are a way of computing how familiar the words in a text are to readers by using the distribution of words and phrases from a large background corpus. We built unigram, bigram, and trigram language models with Good-Turing smoothing over the New York Times section of the English GigaWord corpus (over 900 million words). We used the SRI Language Modeling Toolkit [45] for this purpose. For each of the three n-gram 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.

Word coherence. Word co-occurrence patterns across adjacent sentences provide a way of measuring local coherence which can be easily computed using large amounts of unannotated text [30,44]. Specifically, we used the two features introduced by [44]. [44] make an analogy to machine translation: in translation, two words are likely to be translations of each other if they often appear in *parallel* sentences (a sentence and its translation); in texts, two words are likely to signal local coherence if they often appear in *adjacent* sentences. The two features of word coherence are the *forward likelihood*, the likelihood of observing the words in sentence s_i conditioned on s_{i-1}, and the *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 New York Times section of the English GigaWord corpus. We used the GIZA++[6] implementation of IBM Model 1 to align the words in adjacent sentences and obtain all relevant probabilities.

The equation for the forward likelihood of a text T containing n sentences is below:

$$P_F(T) = \prod_{i=1}^{n-1} \prod_{j=1}^{|s_{i+1}|} \frac{\epsilon}{|s_i| + 1} \sum_{k=0}^{|s_i|} t(s_{i+1}^j | s_i^k) \tag{1}$$

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

Here, sentence s_{i+1} is assumed to be generated from events (words) in sentence s_i. The events in s_i include a special *NULL* word.

The backward likelihood is identical, with s_i and s_{i+1} interchanged.

Entity coherence. Linguistic theories, and Centering theory [20] in particular, have hypothesized that the transition of attention between entities from one sentence to the next plays a major role in the determination of local coherence. [5], inspired by Centering, proposed an easily computable representation for sequences of entity mentions across a text. 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). This representation captures the pattern of entities across sentences in terms of entity transitions. For example, if an entity that occurs in a subject position in sentence s_i is an object in s_{i+1}, the text would have a transition SO. 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] [17] 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.

The actual entity coherence features are the probabilities of local entity transitions (SS, SO, etc), computed as the fraction of each type of transition in the entire entity grid for the text.

4.3 Experimental Setup

We used 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 [14]'s work.

We experiment with the predictive power of the linguistic quality classes of our features 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.

We use a Ranking SVM (SVM^{light} [25]) to learn how to rank summaries using our features. Just as in a SVM used for classification, the Ranking SVM learns a weight vector from the training data. The output of the Ranking SVM is the dot product of the weight vector and the feature values, which is a real number. However, rather than optimizing for this score to be as close as possible to the true score, as in regression, the Ranking SVM instead 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 default regularization parameter was used.

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

Following [5], we report summary ranking accuracy as the fraction of correct pairwise rankings in the test set.

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 thus consists of 51 pairs for human referential clarity; 15,736 pairs for system referential clarity; 57 pairs for human focus; 13,660 pairs for system focus; 88 pairs for human structure; and 14,398 pairs for system structure.

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. Again, we compare all pairs of systems with non-tied gold-standard scores and compute the prediction accuracy for these pairs. At the system level, there are 491 pairs for referential clarity, 492 pairs for focus, and 490 pairs for structure in the test set.

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

4.4 Results

The performance of each class of features is shown in Table 8. The best result in each colum is given in bold, and the rank of the structural features class is noted in brackets.

Structural and language model features are the best predictors of input-level evaluation of human summaries. The pairwise ranking prediction accuracy of structural features is 80% for referential clarity and lower 70s for focus and structure. For system evaluation structural features do reasonably—accuracies of low 60s for input-level and around 85% for system-level for each of the three quality aspects.

No class of predictors stand out as the overall best because the performance differs considerably across tasks. Structural features are very good for input-level human summaries, middle of the range for input level system summaries and about the worst class of features for system-level evaluation of automatic summaries.

Table 8. Pairwise ranking prediction accuracy

Features	**Input-level; Systems**			**Input-level; Humans**			**System-level**		
	Refs	Focus	Struct.	Refs	Focus	Struct.	Refs	Focus	Struct.
LM	62.2	60.5	62.5	76.5	**71.9**	**78.4**	**91.2**	**85.2**	86.3
Coh-metrix	**67.9**	63.0	62.4	68.6	59.6	67.0	88.6	83.9	86.3
Entity coh.	64.3	**64.2**	**63.6**	54.9	52.6	56.8	89.6	85.0	**87.1**
Word coh.	53.3	53.2	53.7	62.7	70.2	60.2	87.8	81.7	79.0
Structural	64.4 [2]	61.9 [3]	62.6 [2]	**80.4 [1]**	**71.9 [1]**	72.7 [2]	87.6 [5]	82.3 [4]	84.9 [4]

The language model and entity coherence classes seem to be the two classes that tend to perform uniformly well for the three tasks.

System-level accuracies are high for all classes of features, above 85% which suggest that using the trained ranker can be a practical substitute of manual evaluation.

5 Conclusion

We presented a study of sentence fluency based on data from machine translation evaluations. These data allow for two types of comparisons: human (fluent) text and (not so good) machine-generated text, and levels of fluency in the automatically produced text. The distinctions were possible even when based solely on features describing syntactic phrasing in the sentences.

Correlation analysis reveals that the structural features are significantly but weakly correlated with fluency. Interestingly, the features correlated with fluency levels in machine-produced text are not the same as those that distinguish between human and machine translations. Such results raise the need for caution when using assessments for machine produced text to build a general model of fluency. The captured phenomena in this case might not be the same as these from comparing human texts with differing fluency. For future research it will be beneficial to build a dedicated corpus in which *human-produced* sentences are assessed for fluency.

Our experiments show that basic fluency distinctions can be made with high accuracy. Machine translations can be distinguished from human translations with accuracy of 87%; machine translations with low fluency can be distinguished from machine translations with high fluency with accuracy of 61%. In pairwise comparison of sentences with different fluency, accuracy of predicting which of the two is better is 90%. Results are not as high but still promising for comparisons in fluency of translations of the same text.

We also demonstrated that while fluency models based on structural features learned on machine translation data do not generalize well to human texts, the models of overall text quality for summarization are robust and can be used for automatic evaluation of linguistic quality. Structural features compare favorably to other classes of predictors of linguistic quality for input-level ranking of human summaries particularly, but also for input-level evaluation of automatic summaries.

References

1. Bailin, A., Grafstein, A.: The linguistic assumptions underlying readability formulae: a critique. Language and Communication 21, 285–301 (2001)
2. Bangalore, S., Rambow, O.: Exploiting a probabilistic hierarchical model for generation. In: Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000), pp. 42–48 (2000)

3. Bangalore, S., Rambow, O., Whittaker, S.: Evaluation metrics for generation. In: Proceedings of the First International Conference on Natural Language Generation (INLG 2000), pp. 1–8 (2000)
4. Banko, M., Mittal, V., Witbrock, M.: Headline generation based on statistical translation. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000), pp. 318–325 (2000)
5. Barzilay, R., Lapata, M.: Modeling local coherence: An entity-based approach. Computational Linguistics 34(1), 1–34 (2008)
6. Barzilay, R., McKeown, K.R.: Sentence fusion for multidocument news summarization. Computational Linguistics 31(3), 297–328 (2005)
7. Cahill, A., Forst, M.: Human evaluation of a German surface realisation ranker. In: Krahmer, E., Theune, M. (eds.) Empirical Methods in NLG. LNCS (LNAI), vol. 5790, pp. 201–221. Springer, Heidelberg (2010)
8. Cahill, A., Forst, M., Rohrer, C.: Stochastic realisation ranking for a free word order language. In: Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 2007), pp. 17–24 (2007)
9. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and maxent discriminative reranking. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 173–180 (2005)
10. Charniak, E.: A maximum-entropy-inspired parser. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics Conference (NAACL 2000), pp. 132–139 (2000)
11. Clarke, J., Lapata, M.: Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006), pp. 377–384 (2006)
12. Collins, M., Koo, T.: Discriminative reranking for natural language parsing. Computational Linguistics 31(1), 25–70 (2005)
13. Collins-Thompson, K., Callan, J.P.: A language modeling approach to predicting reading difficulty. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, pp. 193–200 (2004)
14. Conroy, J., Dang, H.: Mind the gap: dangers of divorcing evaluations of summary content from linguistic quality. In: Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), pp. 145–152 (2008)
15. Corston-Oliver, S., Gamon, M., Brockett, C.: A machine learning approach to the automatic evaluation of machine translation. In: Proceedings of 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001), pp. 148–155 (2001)
16. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41, 391–407 (1990)
17. Elsner, M., Austerweil, J., Charniak, E.: A unified local and global model for discourse coherence. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pp. 436–443 (2007)
18. Galley, M., McKeown, K.: Lexicalized Markov grammars for sentence compression. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pp. 180–187 (2007)

19. Graesser, A., McNamara, D., Louwerse, M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods Instruments and Computers 36(2), 193–202 (2004)
20. Grosz, B., Joshi, A., Weinstein, S.: Centering: a framework for modelling the local coherence of discourse. Computational Linguistics 21(2), 203–226 (1995)
21. Haberlandt, K., Graesser, A.: Component processes in text comprehension and some of their interactions. Journal of Experimental Psychology: General 114(3), 357–374 (1985)
22. Holmes, G., Donkin, A., Witten, I.: Weka: A machine learning workbench. In: Second Australian and New Zealand Conference on Intelligent Information Systems, pp. 357–361 (1994)
23. Huang, L.: Forest reranking: Discriminative parsing with non-local features. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2008: HLT), pp. 586–594 (2008)
24. Jing, H.: Sentence reduction for automatic text summarization. In: Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP 2000), pp. 310–315 (2000)
25. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 133–142 (2002)
26. Just, M., Carpenter, P.: The psychology of reading and language comprehension, Allyn, Bacon (1987)
27. Karamanis, N., Mellish, C., Poesio, M., Oberlander, J.: Evaluating centering for information ordering using corpora. Computational Linguististics 35(1), 29–46 (2009)
28. Knight, K., Marcu, D.: Summarization beyond sentence extraction: a probabilistic approach to sentence compression. Artificial Intelligence 139(1), 91–107 (2002)
29. Langkilde, I., Knight, K.: Generation that exploits corpus-based statistical knowledge. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL 1998), pp. 704–710 (1998)
30. Lapata, M.: Probabilistic text structuring: Experiments with sentence ordering. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), pp. 545–552 (2003)
31. Lapata, M., Barzilay, R.: Automatic evaluation of text coherence: models and representations. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005), pp. 1085–1090 (2005)
32. Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram cooccurrence statistics. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003), pp. 71–78 (2003)
33. Lin, C.: Rouge: A package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), pp. 25–26 (2004)
34. McDonald, R.: Discriminative sentence compression with soft syntactic evidence. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), pp. 297–304 (2006)
35. Mutton, A., Dras, M., Wan, S., Dale, R.: GLEU: Automatic evaluation of sentence-level fluency. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007), pp. 344–351 (2007)
36. Over, P., Dang, H., Harman, D.: DUC in context. Information Processing Management 43(6), 1506–1520 (2007)

37. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 311–318 (2002)
38. Petersen, S.E., Ostendorf, M.: A machine learning approach to reading level assessment. Computer Speech and Language 23(1), 89–106 (2009)
39. Pitler, E., Nenkova, A.: Revisiting readability: a unified framework for predicting text quality. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), pp. 186–195 (2008)
40. Rieser, V., Lemon, O.: Natural language generation as planning under uncertainty for spoken dialogue systems. In: Krahmer, E., Theune, M. (eds.) Empirical Methods in NLG. LNCS (LNAI), vol. 5790, pp. 105–120. Springer, Heidelberg (2010)
41. Schwarm, S., Ostendorf, M.: Reading level assessment using support vector machines and statistical language models. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 523–530 (2005)
42. Siddharthan, A.: Syntactic simplification and Text Cohesion. Ph.D. thesis, University of Cambridge, UK (2003)
43. Soricut, R., Marcu, D.: Abstractive headline generation using WIDL-expressions. Information Processing and Management 43(6), 1536–1548 (2007)
44. Soricut, R., Marcu, D.: Discourse generation using utility-trained coherence models. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pp. 803–810 (2006)
45. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Seventh International Conference on Spoken Language Processing (ICSLP 2002), vol. 3 (2002)
46. Turner, J., Charniak, E.: Supervised and unsupervised learning for sentence compression. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 2005), pp. 290–297 (2005)
47. Velldal, E., Oepen, S.: Maximum entropy models for realization ranking. In: Proceedings of the 10th Machine Translation Summit, pp. 109–116 (2005)
48. Wan, S., Dale, R., Dras, M.: Searching for grammaticality: Propagating dependencies in the Viterbi algorithm. In: Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG 2005), pp. 211–216 (2005)
49. Zajic, D., Dorr, B., Lin, J., Schwartz, R.: Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. Information Processing Management 43(6), 1549–1570 (2007)
50. Zwarts, S., Dras, M.: Choosing the right translation: A syntactically informed classification approach. In: Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), pp. 1153–1160 (2008)