3-2009

Predicting the Fluency of Text with Shallow Structural Features: Case Studies of Machine Tanslation and Human-Written Text

Jieun Chae
University of Pennsylvania

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

Recommended Citation

Jieun Chae and Ani Nenkova, "Predicting the Fluency of Text with Shallow Structural Features: Case Studies of Machine Tanslation and Human-Written Text", . March 2009.

Chae, J. & Nenkova, A., Predicting the Fluency of Text with Shallow Structural Features: Case Studies of Machine Translation and Human-Written Text, *12th Conference of the European Chapter of the Association for Computational Linguistics*, March-April 2009, doi: anthology/E09-1017

Predicting the Fluency of Text with Shallow Structural Features: Case Studies of Machine Tanslation and Human-Written Text

Abstract

Sentence fluency is an important component of overall text readability but few studies in natural language processing have sought to understand the factors that define it. We report the results of an initial study into the predictive power of surface syntactic statistics for the task; we use fluency assessments done for the purpose of evaluating machine translation. We find that these features are weakly but significantly correlated with fluency. Machine and human translations can be distinguished with accuracy over 80%. The performance of pairwise comparison of fluency is also very high—over 90% for a multi-layer perceptron classifier. We also test the hypothesis that the learned models capture general fluency properties applicable to human-written text. The results do not support this hypothesis: prediction accuracy on the new data is only 57%. This finding suggests that developing a dedicated, task-independent corpus of fluency judgments will be beneficial for further investigations of the problem.

Disciplines

Computer Sciences

Comments

Chae, J. & Nenkova, A., Predicting the Fluency of Text with Shallow Structural Features: Case Studies of Machine Translation and Human-Written Text, *12th Conference of the European Chapter of the Association for Computational Linguistics*, March-April 2009, doi: anthology/E09-1017

Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text

Jieun Chae
University of Pennsylvania
chaeji@seas.upenn.edu

Ani Nenkova
University of Pennsylvania
nenkova@seas.upenn.edu

Abstract

Sentence fluency is an important component of overall text readability but few studies in natural language processing have sought to understand the factors that define it. We report the results of an initial study into the predictive power of surface syntactic statistics for the task; we use fluency assessments done for the purpose of evaluating machine translation. We find that these features are weakly but significantly correlated with fluency. Machine and human translations can be distinguished with accuracy over 80%. The performance of pairwise comparison of fluency is also very high—over 90% for a multi-layer perceptron classifier. We also test the hypothesis that the learned models capture general fluency properties applicable to human-written text. The results do not support this hypothesis: prediction accuracy on the new data is only 57%. This finding suggests that developing a dedicated, task-independent corpus of fluency judgments will be beneficial for further investigations of the problem.

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 (Langkilde and Knight, 1998; Bangalore and Rambow, 2000), as well as an important step in machine translation. Considerations of sentence fluency are also key in sentence simplification (Siddharthan, 2003), sentence compression (Jing, 2000; Knight and Marcu, 2002; Clarke and Lapata, 2006; McDonald, 2006; Turner and Charniak, 2005; Galley and McKeown, 2007), text re-generation for summarization (Daumé III and Marcu, 2004; Barzilay and McKeown, 2005; Wan et al., 2005) and headline generation (Banko et al., 2000; Zajic et al., 2007; Soricut and Marcu, 2007).

Despite its importance for these popular applications, the factors contributing to sentence level fluency have not been researched indepth. Much more attention has been devoted to discourse-level constraints on adjacent sentences indicative of coherence and good text flow (Lapata, 2003; Barzilay and Lapata, 2008; Karamanis et al., to appear).

In many applications fluency is assessed in combination with other qualities. For example, in machine translation evaluation, approaches such as BLEU (Papineni et al., 2002) use n-gram overlap comparisons with a model to judge overall "goodness", 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 (Bangalore et al., 2000) (simple accuracy, generation accuracy) are based on string-edit distance from an ideal output.

In contrast, the work of (Wan et al., 2005) and (Mutton et al., 2007) directly sets as a goal the assessment of sentence-level fluency, regardless of content. In (Wan et al., 2005) 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. The idea is extended in (Mutton et al., 2007), where four parsers are used

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 139–147,
Athens, Greece, 30 March – 3 April 2009. ©2009 Association for Computational Linguistics

and artificially generated sentences with varying level of fluency are evaluated with impressive success. The fluency models hold promise for actual improvements in machine translation output quality (Zwarts and Dras, 2008). In that work, only simple parser features are used for the prediction of fluency, but no actual syntactic properties of the sentences. But certainly, problems with sentence fluency are expected to be manifested in syntax. We would expect for example that syntactic tree features that capture common parse configurations and that are used in discriminative parsing (Collins and Koo, 2005; Charniak and Johnson, 2005; Huang, 2008) should be useful 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 (Corston-Oliver et al., 2001). 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 mismatch 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.

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 revisit the task of distinguishing machine translations from human translations, but also further our understanding of fluency by providing comprehensive analysis of the association between fluency assessments of translations and surface syntactic features. We also demonstrate that based on the same class of features, it is possible to distinguish fluent machine translations from disfluent machine translations. 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.

For our experiments we use the evaluations of Chinese to English translations distributed by LDC (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.
- Use the models trained on data from MT evaluations to predict potential fluency problems of human-written texts (from the Wall Street Journal).

Even for the last most challenging task results are promising, with prediction accuracy almost 10% better than a random baseline. For the other tasks accuracies are high, exceeding 80%.

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. No general corpora exist with fluency assessments, 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 Features

Perceived sentence fluency is influenced by many factors. The way the sentence fits in the context of surrounding sentences is one obvious factor (Barzilay and Lapata, 2008). Another well-known factor is vocabulary use: the presence of uncommon difficult words are known to pose problems to readers and to render text less readable (Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005). But these discourse- and vocabulary-level features measure properties at granularities different from the sentence level.

Syntactic 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 discuss discourse and language model features that have been extensively studied in prior work on coherence and readability.

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 (Charniak, 2000) in order to calculate these features.

Sentence length is the number of words in a sentence. Evaluation metrics such as BLEU (Papineni et al., 2002) 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 is considered to be a measure of sentence complexity. Generally, longer sentences are syntactically more complex but when sentences are approximately the same length the larger 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 Out of the 2634 total sentences, only 165 contained a fragment tag in their parse, indicating the presence of ungrammaticality in the sentence. Fragments occur 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—one with embedded phrases included in the calculation and one just for the largest phrases of a given type. *Normalized average phrase length* 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 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 longest phrase (where embedded phrases of the same type were counted only once) and for each phrase regardless of embedding.

Length of NPs/PPs contained in a VP The average number of words that constitute an 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 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.

3 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.

Pearson's correlation between the fluency ratings and features are shown in Table 2. First of all, fluency and adequacy as given by MT evaluators

Fluency score	The 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 1: Distribution of fluency scores.

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. 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 signal potential fluency problems:

- 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 (Bailin and Grafstein, 2001). 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.

4 Experiments with machine translation data

4.1 Distinguishing human from machine translations

In this section we use all the features discussed in Section 2 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 Weka—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 fluency scores vs. lowest fluency score). We expect that low fluency MT will be more easily

adequacy 0.701(0.00)	**sentence length** -0.132(0.00)	**unnormalized NP length** -0.124(0.00)	**VP distance** -0.116(0.00)
unnormalized VP length -0.109(0.00)	**Max Tree depth** -0.106(0.00)	**phrase length** -0.105(0.00)	**avr. NP length (embedded)** -0.097(0.00)
avr. VP length (embedded) -0.094(0.00)	**SBAR length** -0.086(0.00)	**avr. largest NP length** -0.084(0.00)	**Unnormalized PP** -0.082(0.00)
avr PP length (embedded) -0.070(0.00)	**SBAR count** -0.069(0.001)	**PP length in VP** -0.066(0.001)	**Normalized PP1** 0.065(0.001)
NP length in VP -0.058(0.003)	**PP length** -0.054(0.006)	**normalized VP length** 0.054(0.005)	**PP length in NP** 0.053(0.006)
Fragment -0.049(0.011)	**avr. ADJP length (embedded)** -0.046(0.019)	**avr. largest VP length** -0.038(0.052)	

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

	worst 300 MT	**best 300 MT**	**total MT (5920)**
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%

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

distinguished from human translation in comparison with machine translations rated as having high fluency.

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 results 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%.

4.2 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. For every two sentences, the feature for the pair is 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.

In fact, the problem is *much* more difficult as

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%

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 data set.

can be seen in the second row of Table 4. 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 impressively high, and significantly higher than baseline performance. The results are about 20% lower than for prediction of a more fluent sentence when the task is not constrained to translation of the same sentence.

4.3 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 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.

5 Applications to human written text

5.1 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 (Pitler and Nenkova, 2008). 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 machine MT), the other was the model for distinguishing the 300 best from the 300 worst machine translations (good vs bad MT). The classifiers used were decision trees for human vs machine distinction and support vector machines for good 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 "best MT" class.

The results are shown in Table 6. The two models vastly differ in performance. 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

MT vs HT	good MT vs Bad MT	Ranking	Same sentence Ranking
unnormalized PP	SBAR count	avr. NP lengt	normalized NP length
PP length in VP	Unnormalized VP length	normalized PP length	PP count
avr. NP length	post attribute length	NP count	normalized NP length
# apposition	VP count	normalized NP length	max tree depth
SBAR length	sentence length	normalized VP length	avr. phrase length

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

Model	Acc	P	R
human vs machine trans.	57%	0.79	0.58
good MT vs bad MT	44%	0.57	0.44

Table 6: Accuracy, precision and recall (for fluent class) for each model when test on WSJ sentences. The gold-standard is assessment by a single reader of the text.

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.

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

Model and assessor agree that sentence is problematic:

(1.1) 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.

(1.2) 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.

(1.3) 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.

The model predicts the sentence is good, but the assessor finds it problematic:

(2.1) 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.

(2.2) 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%.

(2.3) 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.

The model predicts the sentences are bad, but the assessor considered them fluent:

(3.1) The sense grows that modern public bureaucracies simply don't perform their assigned functions well.

(3.2) Amstrad PLC, a British maker of computer hardware and communications equipment, posted a 52% plunge in pretax profit for the latest year.

(3.3) At current allocations, that means EPA will be spending $300 billion on itself.

5.2 Correlation with overall text quality

In our final experiment 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 assess the overall text quality of each article on a scale from 1 to 5 (Pitler and Nenkova, 2008). The average of the assessments was taken as a single number describing the 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. The positive correlation would suggest that the more hard to read sentence appear in a text, the higher the text would be rated overall, which is surprising. The predictions from the model for distinguishing good and bad machine translations very close to zero, but negative which corresponds better to the intuitive relationship between the two.

Note that none of the correlations are actually significant for the small dataset of 30 points.

6 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

Fluency given by	Correlation
human	0.127
human vs machine trans. model	-0.055
good MT vs bad MT model	0.076

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

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 significant 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 be different than 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. The prediction becomes better when the texts being compared exhibit larger difference in fluency quality.

Admittedly, our pilot experiments with human assessment of text quality and sentence level fluency are small, so no big generalizations can be made. Still, they allow some useful observations that can guide future work. They do show that for further research in automatic recognition of fluency, new annotated corpora developed specially for the task will be necessary. They also give some evidence that sentence-level fluency is only weakly correlated with overall text quality. Discourse apects and language model features that have been extensively studied in prior work are indeed much more indicative of overall text quality (Pitler and Nenkova, 2008). We leave direct comparison for future work.

References

A. Bailin and A. Grafstein. 2001. The linguistic assumptions underlying readability formulae: a critique. *Language and Communication*, 21:285–301.

S. Bangalore and O. Rambow. 2000. Exploiting a probabilistic hierarchical model for generation. In *COLING*, pages 42–48.

S. Bangalore, O. Rambow, and S. Whittaker. 2000. Evaluation metrics for generation. In *INLG'00: Proceedings of the first international conference on Natural language generation*, pages 1–8.

M. Banko, V. Mittal, and M. Witbrock. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the Association for Co mputational Linguistics*.

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

R. Barzilay and K. McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3).

E. Charniak and M. Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *NAACL-2000*.

J. Clarke and M. Lapata. 2006. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *ACL:COLING'06*, pages 377–384.

M. Collins and T. Koo. 2005. Discriminative reranking for natural language parsing. *Comput. Linguist.*, 31(1):25–70.

K. Collins-Thompson and J. Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL'04*.

S. Corston-Oliver, M. Gamon, and C. Brockett. 2001. A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 148–155.

H. Daumé III and D. Marcu. 2004. Generic sentence fusion is an ill-defined summarization task. In *Proceedings of the Text Summarization Branches Out Workshop at ACL*.

M. Galley and K. McKeown. 2007. Lexicalized markov grammars for sentence compression. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Liang Huang. 2008. Forest reranking: Discriminative parsing with non-local features. In *Proceedings of ACL-08: HLT*, pages 586–594.

H. Jing. 2000. Sentence simplification in automatic text summarization. In *Proceedings of the 6th Applied NLP Conference, ANLP'2000*.

N. Karamanis, M. Poesio, C. Mellish, and J. Oberlander. (to appear). Evaluating centering for information ordering using corpora. *Computational Linguistics*.

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

I. Langkilde and K. Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *COLING-ACL*, pages 704–710.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL'03*.

R. McDonald. 2006. Discriminative sentence compression with soft syntactic evidence. In *EACL'06*.

A. Mutton, M. Dras, S. Wan, and R. Dale. 2007. Gleu: Automatic evaluation of sentence-level fluency. In *ACL'07*, pages 344–351.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

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

S. Schwarm and M. Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of ACL'05*, pages 523–530.

A. Siddharthan. 2003. *Syntactic simplification and Text Cohesion*. Ph.D. thesis, University of Cambridge, UK.

R. Soricut and D. Marcu. 2007. Abstractive headline generation using widl-expressions. *Inf. Process. Manage.*, 43(6):1536–1548.

J. Turner and E. Charniak. 2005. Supervised and unsupervised learning for sentence compression. In *ACL'05*.

S. Wan, R. Dale, and M. Dras. 2005. Searching for grammaticality: Propagating dependencies in the viterbi algorithm. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

D. Zajic, B. Dorr, J. Lin, and R. Schwartz. 2007. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Inf. Process. Manage.*, 43(6):1549–1570.

S. Zwarts and M. Dras. 2008. Choosing the right translation: A syntactically informed classification approach. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1153–1160.