Treat the system like a human student: Automatic naturalness evaluation of generated text without reference texts

Ye Tian **Ioannis Douratsos** **Isabel Groves**
Amazon Research Cambridge
Cambridge, UK
{yetiancl, ioannisd, isabeg}@amazon.co.uk

Abstract

The current most popular method for automatic Natural Language Generation (NLG) evaluation is comparing generated text with human-written reference sentences using a metrics system, which has drawbacks around reliability and scalability. We draw inspiration from second language (L2) assessment and extract a set of linguistic features to predict human judgments of sentence naturalness. Our experiment using a small dataset showed that the feature-based approach yields promising results, with the added potential of providing interpretability into the source of the problems.

1 Introduction

More and more text is generated in Machine Translation, Text Summarization, Image Captioning, and Dialogue Systems. With this increased usage of Natural Language Generation (NLG) comes an increase in the importance of evaluating the language generated, and an increase in the difficulty of doing so as the quantity and variety of output increases. Automatic NLG evaluation focuses on two areas: accuracy and fluency. The former assesses how well the generated text conveys the desired meaning, while the latter assesses how well the language flows: the 'linguistic quality of the text' (Gatt and Krahmer, 2018) and whether it sounds like something a native speaker of the language would naturally produce. This paper focuses on the latter. We first review current approaches in metrics-based evaluation, in referenceless evaluation and in second language (L2) language assessment; we then present our experiment in section 3.

1.1 Metrics system using human reference set - the lion's share

NLG evaluation has traditionally relied on human judgments (Mellish and Dale, 1998). Beyond that, the predominant automated method is to compare generated text with one or more human-created reference texts using a metric-based system (Gatt and Krahmer, 2018). The more similar the system output is to the human authored text, the better the system is judged to be. Popular metrics include BLEU (Papineni et al., 2002), ROUGE (Lin and Hovy, 2003), NIST (Doddington, 2002), METEOR (Lavie and Agarwal, 2007) and CIDEr (Vedantam et al., 2015), among others. Up to 60% of NLG research published between 2012 and 2015 relied on such metrics (Gkatzia and Mahamood, 2015)

However, it has repeatedly been found that automated metrics do not correlate well with human evaluations of generated text (Stent et al., 2005; Belz and Reiter, 2006; Reiter and Belz, 2009) and that the correlation is weaker at sentence-level than when evaluating a system overall. (Novikova et al., 2017a; Shimorina, 2018). Novikova et al. (2017a) compared popular comparison metrics used to evaluate NLG systems, concluding that the current state-of-the-art metrics are insufficient and cannot replace human judgments. They demonstrated that all the aforementioned automated metrics based on word-overlap with reference texts were strongly correlated with each other and only weakly correlated with human judgments of naturalness and quality. Furthermore, the least weak correlation found between any metric and human naturalness judgments was on the least varied dataset that only expressed a limited set of attributes and had less lexical diversity as it was only partially lexicalised (all proper names were replaced by placeholder variables). Given that lex-

Proceedings of The 11th International Natural Language Generation Conference, pages 109–118,
Tilburg, The Netherlands, November 5-8, 2018.

icalisation is a source of ungrammaticality in NLG (Sharma et al., 2016), this dataset therefore does not fully represent the challenge of evaluating the final output of an NLG system.

In addition to accuracy concerns, using a metrics system with a human reference set has several practical limitations. Firstly, building reference sets tends to require experts (e.g. translators) and is thus costly to create. Secondly, an output that is different from a human-written reference is not necessarily a bad sentence for the task: there are often multiple valid ways to express a desired meaning. The evaluation therefore requires multiple reference sentences, which makes producing a reference set even harder and generates complexities in similarity calculation. Thirdly, creating a human gold standard is not suitable for fast or large scale assessment. For NLG systems that cover a large variety of topics, the quantity of reference sentences required can be prohibitive to using this approach during system development.

1.2 Moving away from human reference set

We should look beyond evaluation using human references and learn from research outside our immediate domain, since there has been more research into automatic evaluation of text without human references in tasks similar to NLG than there has been for NLG itself.

One such domain is second language learner (L2) language assessment. Here the target is not machine-generated text but human-produced text. Over the last decade, a large body of work has identified linguistic features that indicate language fluency and complexity (Hancke et al., 2012; Feng, 2010; Chen and Zechner, 2011; Lu, 2010; Vajjala and Meurers, 2012). The linguistic feature based models in L2 assessment seem to correlate more strongly with human judgments of naturalness than current NLG evaluation metrics (with the caveat that these are different tasks). Many of the features require syntactic and discourse parsing, and they capture linguistic knowledge of what makes sentences readable and natural, as reflected in psycholinguistic studies on reading and parsing effort. These features are often more interpretable than purely statistical metrics, so potentially they allow us to not only evaluate the naturalness of a sentence or document, but also to identify *why* it is good or bad.

Another relevant domain is automatic grammaticality judgment. Wagner et al. (2009) investigated grammaticality classification using features such as part-of-speech (POS) n-gram frequencies and the output of probabilistic parsers trained on corpora of grammatical and ungrammatical sentences. They found that parse probability is reduced by spelling, agreement and verb form errors. Heilman et al. (2014) also found linguistic feature based models to be effective when using spelling, language model and grammar features from different parsers. They found that n-gram frequencies and the ability to be parsed were the most influential features for indicating grammaticality. This feature-based method also proved effective in grammaticality evaluation when applied to grammatical error correction applications (Napoles et al., 2016).

In Machine Translation, quality estimation without reference texts has been the subject of multiple shared tasks (Bojar et al., 2017). The QuEst 2015 sentence level model (Specia et al., 2015)[1] that provided the baseline for the latest completed task uses features of the source and/or target sentences including features from language model scores, length, part-of-speech and dependency parsing The leading system (Kim et al., 2017) in the 2017 task used an end-to-end stacked neural model consisting of a bilingual neural word prediction model and neural quality estimator model. The next best performing team's submission (Martins et al., 2017) used a stacked combination of a linear feature-based model (with dependency, POS and syntactic features) with a neural network.

Within NLG evaluation, Novikova et al. (2017a) examined the correlation between human evaluations and grammar-based measures that indicate readability and grammaticality. To measure grammaticality, they used the number of misspellings and the Stanford parser parsing score. Using the Flesch Reading Ease score (Flesch, 1979) and various other measures of complexity such as character, word, syllable and sentence counts, they found that, at a system level, systems producing utterances of higher readability and shorter word length received higher naturalness and overall quality ratings from humans. However, at sentence level there was no strong correlation between such metrics and human ratings that could reliably identify generated sen-

[1] http://www.quest.dcs.shef.ac.uk

tences with low readability or low grammaticality. This evidence that the linguistic features of texts do correlate with human judgments in NLG but that no single feature does so with a strong correlation supports our proposal that combining multiple grammatical features could automatically identify the quality of generated sentences.

We apply the feature-based approach used elsewhere by trying to identify whether machine-generated sentences are fluent and natural, and compare the predictions with human produced labels. Unlike previous work on grammaticality prediction we focus on the notion of "naturalness" or "fluency" rather than just grammaticality. This is because 1) psycholinguistic studies have shown that human perception of grammaticality is gradient (Keller, 2001), and 2) for most systems involving NLG, it matters how easy it is for humans to understand the sentences, not just whether the sentences are grammatical. With this in mind, we use features to capture the ease of parsing (influenced by grammaticality and syntactic complexity) and semantic soundness (influenced by word collocations and frequency). One recent investigation into NLG evaluation without reference texts that we are aware of used a recurrent neural network to estimate quality using the meaning representation input and output sentence to estimate the overall quality (Dušek et al., 2017). Our work differs in the use of linguistic features, which have proved successful in other domains and offer the prospect of interpretability, and we maintain the separation between evaluating the adequacy of the semantic content and evaluating the fluency of the text as has been found to be advisable for NLG evaluation (Stent et al., 2005).

2 Deriving the linguistic feature set

Expanding on the literature on L2 language assessment, especially (Hancke et al., 2012), and on grammaticality evaluation, we derived five groups of features (see full list in Table 1).

2.1 Lexical features

Lexical features include counts and ratios of words, lemmas and Part-of-Speech (POS) tokens. Type-Token Ratio (TTR), the ratio of the number of word types (in terms of lemmas) to total number of word tokens in a text, and its variants are used to measure lexical variation in language acquisition studies. We adopted the variations described in (Vajjala and Meurers, 2012) and word counts by POS categories, extracted using spaCy[3].

2.2 Constituency parse features

We used the BLLIP reranking parser (Charniak and Johnson, 2005), which includes a generative constituent parser and a discriminative maximum entropy reranker, and the WSJ-Gigaword-v2 model which consists of the Wall Street Journal corpus from Penn Tree Bank and two million sentences from Gigaword. From the parser output we used as features the parser log probability and reranker log probability of the most likely parse after reranking the 50-best parses. The idea is that parse probability reflects parser confidence and correlates with sentence quality (Mutton et al., 2007). We also added features for kurtosis and skew of the log probabilities of the 50 most likely parses, based on the idea that the distribution reflects sentence grammaticality and readability (Wagner et al., 2006). Our intuition was that a well-formed grammatical sentence would have positive skew and high kurtosis dropping steeply from the highly probable best parse to other much less likely parses. Conversely, an ungrammatical sentence would have a flatter kurtosis as none of the parses are very probable. Other features include tree height (length of the longest path from the root), number of subtrees, proportion of non-terminal subtrees, the number and mean token length of Noun Phrase (NP), Verb Phrase (VP) and Adjective Phrase (AdjP) sub-trees.

2.3 Dependency parse features

Using the spaCy dependency parser, we extracted the root word of the dependency tree and its part of speech, the tree height and the subtree height to either side of the root. The part of speech of the root is an indicator of whether the sentence has a main verb. The size of the tree on either side of the root reflects whether a sentence is "top" or "tail" heavy, or more balanced. This feature is based on the principle that sentences are easier to process, and thus are judged to be natural and well worded, if the dependencies of the head are roughly evenly distributed on either side (Temperley, 2008), and that heavy noun phrases are hard to process at the beginning of the sentence (Stallings et al., 1998).

[3]https://github.com/explosion/spaCy

Lexical Features		**Constituency Parse Features**	
Type-Token Ratio(TTR)	Num nouns	Constituency Tree height	Num NPs
Root TTR*	Num verbs	Parser probability*	NP average length
Corrected TTR*	Num possessives	Reranker probability*	Num VPs
Bilogarithmic TTR	Num preposition	50-best reranker score kurtosis*	VP average length*
Uber Index	Num determiners	50-best reranker score skew*	Num PPs
Lexical Density	Num adjectives	Num subtrees	PP average length
Answer length	Num relative pronouns	Num non-terminal subtrees	
Lexical repetition*	Num digits	% of non-terminal subtrees	
Num tokens	Num conjunctions		
Dependency Parse Features		**Language Model Features**	
Dependency tree height	Left subtree height	POS LM - Unigram	POS LM - Bigram*
Right subtree height	Num words left of root	POS LM - Trigram	Words LM - Score*
Num words right of root	Root POS	Words LM - Perplexity*	
Grammar Checker	LanguageTool		

Table 1: Feature list. Highest contribution features indicated by *

	class "Not Perfect"		**class "Perfect"**		Weighted	Overall
Model	Precision	Recall	Precision	Recall	F1	Accuracy
Baselines						
Baseline always predicting "Not Perfect"	.84	1	0	0	.76	.84
Deep Learning Baseline	.85	.97	.42	.12	.79	.83
Feature-based models						
Random Forest	.90	.97	.77	.45	.88	.89
Logistic Regression[2]	.91	.96	.70	.49	.87	.88
Feature ablation						
LM perplexity only - KNeighbors	.84	.1	.60	.02	.77	.84
Parser reranker probability only - KNeighbors	.87	.97	.63	.27	.86	.83
Top 11 ranked features - Random Forest	.90	.97	.75	.46	.88	.89

Table 2: Results of baselines, top two feature-based classifiers and models using subset of features.

2.4 Language Model based features

A Language Model (LM) represents the probability distribution of n-grams in a corpus and can measure how "surprised" the model is to see a sentence. We used both POS-based LMs and word-based LMs. For POS-LMs, the POS sequences of each sentence were evaluated against unigram, bigram and trigram POS-based LMs trained on the Wall Street Journal corpus made available in CoNLL2000 (Tjong Kim Sang and Buchholz, 2000). Word-based LMs were trained using the KenLM package (Heafield et al., 2013). We trained two models, one using an English news corpus (available at (Heafield et al., 2013)), and the other using WikiText (Merity et al., 2016). The score was calculated as $\log_{10} p(sentence\langle/s\rangle|\langle s\rangle)$ where $\langle s\rangle$ and $\langle/s\rangle$ are the symbols for beginning and end of sentence, respectively. This reflects, after seeing a start-of-sentence symbol, the probability of a sentence appearing and being followed by an end-of-sentence token. Perplexity of a sentence was calculated with $10.0^{\frac{-score(sentence)}{length(words)+1}}$.

2.5 Grammar checker

We used the open source rule-based grammar checker LanguageTool[4] (Naber, 2003) to output a binary label of whether a sentence violates any of the English grammatical rules encoded in this tool.

3 Experiment

3.1 Data description

We collected our ground-truth evaluations through Amazon Mechanical Turk, asking participants to read machine-generated sentences and judge whether or not they are "perfectly good" English sentences. We opted for a binary judgment task rather than a graded one to make the judgment task simple for participants. The sentences evaluated were 4000 machine-generated sentences from the data released in the 2007/2008 Workshops on Statistical Machine Translation[5]. We did not use the provided human evaluation results because these were evaluations of adequacy, i.e. a mixture of overall quality, content accuracy, and fluency, and the labels were system rankings. We

[4]https://languagetool.org/, "Grammar" category only.

[5]http://www.statmt.org/wmt08/shared-evaluation-task.html

randomly allocated 4000 generated sentences into 40 lists. Each participant read 100 sentences and judged whether each was a "perfectly good" sentence that would sound grammatical and natural to someone with a high proficiency in English. Each sentence was judged by at least 5 participants. Overall, most sentences received the "Not Perfect" rating (Figure 1). The Fleiss kappa on the whole data set is 0.3. We then categorized sentences into "Perfect" (more than 70% "Perfect" judgments), "Not Perfect" (less than 30% "Perfect" judgments), and "Not Sure" (the remainder). There were 603 "perfect" sentences and 2637 "Not Perfect" ones, which were used for model training and evaluation. The 929 "not sure" sentences were excluded.

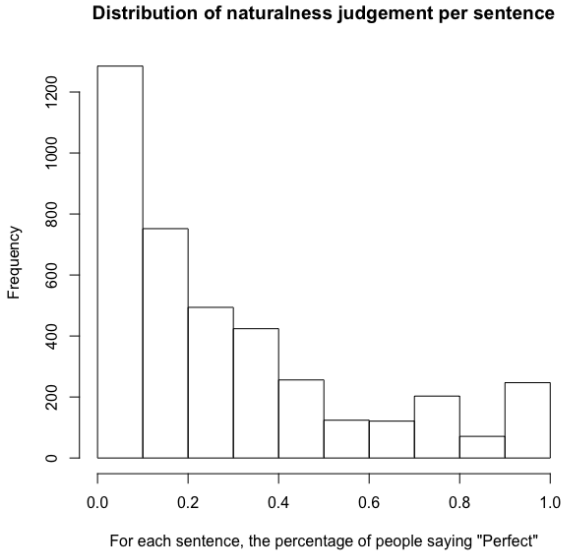

Figure 1: Percentage of "perfect" judgments per sentence

3.2 Training a classifier: Results

We trained "naturalness" classifiers in two ways: using a deep learning model on sentences represented by FastText word embeddings (Bojanowski et al., 2017), and using linguistic features. The deep learning model uses a pooled bidirectional Gated Recurrent Unit (GRU) architecture (Chung et al., 2014). After excluding data with missing feature values, there were 2934 observations for the models, 512 of which were "perfect". We split the data into three sets of equal size, two for training and one for testing.

Given the small dataset, the deep learning model serves as a baseline. It attained a marginally better weighted F1 than an "assume-all-not-perfect" baseline and a similar accuracy.

For the feature based models, we scaled numerical features to be centered around 0 with a standard deviation of 1. Categorical features were encoded in an 1-hot fashion so each level becomes a feature on its own. Using Scikit-learn (Pedregosa et al., 2011), we trained the following classifiers: Linear LVC with L1, L2 or combined penalty, Logistic Regression, KNeighbours Classifier, RandomForest, Perceptron, SGDClassifier and XGboost (Chen and Guestrin, 2016). We used the optimal hyper-parameters for each classifier acquired after running a 5-fold cross validation. We trained all classifiers 10 times and calculated the mean accuracy and F1 of the 10 sessions. The top six classifiers had very similar performances (Logistic Regression, LinearLVC with L1, L2 or combined penalty, RandomForest, SGD classifier). We report the mean results of the top two models in Table 2.

3.3 Error Analysis

When predicting the naturalness of 969 sentences, of which 158 were " Perfect", the top performing RandomForest model labeled 861 out of 969 (88.85%) correctly. It produced 87 incorrect "Not Perfect" labels, and 21 incorrect "Perfect" labels. The incorrect "Not Perfect" labels consisted of three main categories: long sentences (especially those with subordinate clauses), split sentences with inserts (e.g. "I shall, *of course*, inform the President of your comment.") and non-sentential segments that human judges deemed natural (e.g. "The Value of European Values."). Among the incorrect "Perfect" labels, some were assigned to sentences with isolated grammatical errors, such as incorrect verb agreement (e.g. "The Nobel laureate Gary Becker *disagree* with this view."), incorrect prepositions (e.g. "The journal Science *on* the issue last autumn published several contributions.", or word order errors (e.g. "What *now we can* do?"). The overall impression is that the sentences judged to be "Perfect" by the model are easier to read, and are less complex than ones judged to be "Not Perfect".

3.4 Feature Analysis

Different classifiers agreed on the top weighted features, but gave different rankings to features with lighter weight. The highest ranking feature for the top six classifiers is the parser-reranker probability, echoing previous findings that parse probability can be used to evaluate grammaticality

(Mutton et al., 2007). Other top features include number of tokens, number of verbs, constituency tree height and dependency tree height. The effectiveness of Language Model Perplexity and Score is sensitive to the corpora that the model is trained on. In this experiment, LM features trained on the Wikipedia data gave the whole model a .02% boost in F1 compared to LM scores trained on news corpora. We also tested a classifier that used the language model perplexity as the only feature in training and testing, and found this to be less accurate. This indicates that although a language model captures some notion of the likelihood of a sentence, it does not fully encapsulate all that is involved in making a sentence sound natural. Perhaps surprisingly, LanguageTool contributed very little. We realized that the rules it uses to detect grammatical errors are mostly linear and struggle with constituents involving longer dependencies. For example, LanguageTool judged the sentence "I represent a number of sugar beet growers and I am therefore very concerned." to violate the rule "MANY_NN_U", meaning that the quantifier "a number of" is followed by the uncountable noun "sugar", while the actual head noun is "growers".

For a feature ablation study, we used the Scikit-learn implementation of Recursive Feature Elimination to identify which features contributed most to the best performing model, the Random Forest Model. Retraining and testing on subsets of features found that using just the 11 best-performing features achieves the same F1 and accuracy as the model that used all the features. Adding additional lower-ranked features beyond that brought no significant additional benefit (Figure 2). These 11 features were: parser probability, reranker probability, reranker score kurtosis, reranker score skew, average length of verb phrases, the POS language model bigram score, root TTR, corrected TTR, lexical repetition, language model score and language model perplexity.

4 Model and Feature Set transferability

How well would our naturalness model trained on a small dataset in one domain - MT generated sentences about European politics - perform on an entirely different domain? To test the transferability, we used data provided by Novikova et al. (2017a)[6] of sentences produced by NLG systems participating in an end-to-end (E2E) NLG challenge[7] (Novikova et al., 2017b). We used the data from the lexicalised datasets SFRES and SFHOT datasets and the system outputs from the LOLS (Lampouras and Vlachos, 2016)[8] and RNNLG (Wen et al., 2015)[9] NLG systems. These sentences describe restaurant types, locations and categories to convey information given in a *slot+value* meaning representation. This provided 1954 unique sentences. We used the annotations for naturalness that human evaluators had provided on a 6-point Likert scale in response to the question '*Could the utterance have been produced by a native speaker?*'. For each unique system-generated response we took the mean naturalness score across the different annotators. As our model was trained for the task of identifying data as "perfect" versus "imperfect", we set a high threshold for naturalness: responses with a mean naturalness rating of greater than or equal to 5 and no single naturalness score below 5 were set with a ground-truth of perfect. This resulted in 426 "perfect" targets out of 1954 sentences. Using the model described above to predict the naturalness of this dataset resulted in an accuracy of .70 and a weighted F1 of .69. As a baseline for this dataset, always predicting 'imperfect' would have an accuracy of .78 and a weighted F1 of .68. Additionally, we used our classifier training and testing pipeline on this dataset, training on two thirds of the data (1309 sentences) and testing on the other third (645 sentences, of which 126 were 'perfect'). This surpassed the baseline for this dataset: across ten repetitions the mean weighted F1 was .73 and accuracy was .83. Repeating the exercise with just the top 11 features identified during the Feature Analysis above also

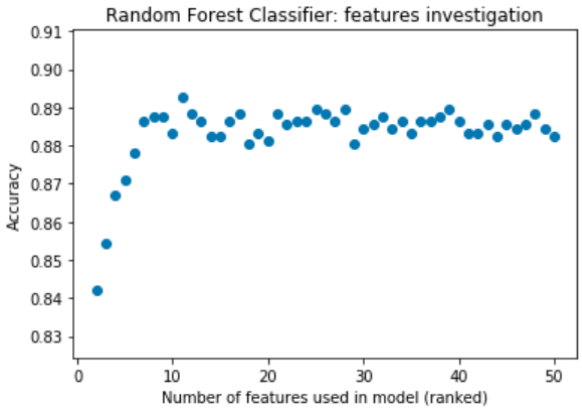

Figure 2: Accuracy results of Random Forest models using a subset of features, ranked by Recursive Feature Elimination

[6]https://github.com/jeknov/EMNLP_17_submission

[7]http://www.macs.hw.ac.uk/InteractionLab/E2E/

[8]https://github.com/glampouras/JLOLS_NLG

[9]https://github.com/shawnwun/RNNLG

Metric	Correlation with mean naturalness	p value
Our model	**0.23**	$p < 0.001$
METEOR	0.18	p <0.001
ROUGE L	0.17	p <0.001
Bleu 2	0.16	p <0.001
Bleu 1	0.15	p <0.001
CIDEr	0.15	p <0.001
Bleu 3	0.15	p <0.001
NIST	0.11	p <0.01
Bleu 4	0.11	p <0.01

Table 3: E2E NLG Challenge data: Spearman's ρ for mean fluency and grammaticality human judgments (model trained on E2E task data).

surpassed the baseline though was lower than the full feature set, resulting in a mean weighted F1 of .73 and an accuracy of .80. (always predicting 'imperfect' would achieve an F1 of .72 and accuracy of .80)

The model's predictions for this test set correlated weakly with the mean naturalness score with a Spearman's ρ of 0.23 ($p < 0.001$) (Table 3). Though this correlation is not very strong, it is notable that it is stronger than the correlation with all the other word-overlap metrics investigated by (Novikova et al., 2017a) and does not require a reference text to achieve this.

We also tested transferabiltiy with data from the WebNLG challenge[10] (Gardent et al., 2017) in order to test on more diverse content about different topics. The WebNLG data consists of sets of triples extracted from DBPedia across 15 different categories carefully designed to be varied. Utterances generated by WebNLG Challenge entrants underwent human annotation by participants from English-speaking countries. We used the annotations for fluency and grammaticality[11] which were graded separately, each on a three-point Likert scale. We set the ground truth of 'perfect' for those sentences which had a mean fluency and grammaticality annotation greater than or equal 2.6 with no single annotation lower than 2. This gave us 1959 unique sentences of which 624 were 'perfect'. Our original model's predictions resulted in an accuracy of 0.68 and a weighted F1 of 0.61. A baseline for this dataset that always predicted 'imperfect' would have an accuracy of 0.78 and an F1 of 0.55. As with the E2E set, performance

[10]http://webnlg.loria.fr/pages/challenge.html

[11]https://gitlab.com/shimorina/webnlg-human-evaluation/

	Correlation with fluency	Correlation with grammaticality
Our model	0.35	**0.46**
Bleu	0.33	0.28

Table 4: WebNLG Challenge data: Spearman's ρ correlation with mean fluency and grammaticality human judgments (model trained on WebNLG task data). All p <0.001

improved when trained on data from this task. We used our pipeline to train a model on this data, split two thirds/one third between training and testing giving a test set of 647 of which 433 were 'perfect'. This resulted in an accuracy of 0.71 and a weighted F1 of 0.69 (the mean over 10 iterations). A baseline for this test set that always predicted 'imperfect' would have an accuracy of 0.44 and an F1 of 0.55. This indicates that our feature set can capture some characteristics of what constitutes a well-worded response in these domains also.

We use the Bleu scores that had been calculated using the dataset's reference sentences to compare Bleu's correlation with fluency and grammaticality judgments and the correlation with our model's predictions. The original model correlates very weakly with mean fluency score (Spearman's ρ 0.08, p <0.001) and does not correlate significantly with mean grammaticality score p >0.05). However, when trained on this task, the model's predictions were moderately and significantly positively correlated with the mean fluency and grammaticality ratings (Table 4). The correlation with Bleu is weaker on this test set: trained on data from this task, we achieve better correlation with fluency and in particular grammaticality judgments than Bleu.

This exercise shows that while our model may have limited direct transferability when there are significant differences between the type of sentences seen in the training data domain versus the test, our feature-based method and feature set are more transferable than the model itself. When trained on data for a different task, different features from the set can contribute to identifying what constitutes a high quality sentence in this genre. This approach could be used to evaluate the naturalness of generated text for a particular task by using a small set of human-annotated data to train a model that can cheaply and easily be used over a larger quantity of data to given an indication of the naturalness.

5 Conclusions and Future Work

We presented a linguistic feature based approach to automatic naturalness evaluation of machine generated text, building on findings from L2 assessment research. Our experiment using a small dataset showed promising results suggesting that this is a viable path towards scalable naturalness evaluation of machine-generated text, with potential for interpretability which can help identify and prioritize improvements to an NLG system during development. In future work, we aim to extend this approach to outputs in multiple languages and multiple domains to further assess the transferability of the approach and of specific models. We will go beyond a binary classification of "perfect" versus "imperfect" to better account for cases where there is inter-speaker variation in naturalness judgments. We also plan to investigate improving deep neural models by adopting recent advancements in contextualized deep word and sentence embeddings (Peters et al., 2018; Perone et al., 2018) and transfer learning in sentence representation (Howard and Ruder, 2018; Radford et al., 2018).

References

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics* 5(1):135–146.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*. pages 169–214.

Eugene Charniak and Mark 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)*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 173–180. http://www.aclweb.org/anthology/P/P05/P05-1022.pdf.

Miao Chen and Klaus Zechner. 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 722–731.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, pages 785–794.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* abs/1412.3555. http://arxiv.org/abs/1412.3555.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, HLT '02, pages 138–145. http://dl.acm.org/citation.cfm?id=1289189.1289273.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2017. Referenceless quality estimation for natural language generation. *Proceedings of the 1st Workshop on Learning to Generate Natural Language, Sydney, Australia.* .

Lijun Feng. 2010. *Automatic readability assessment*.

Rudolf Franz Flesch. 1979. *How to write plain English: A book for lawyers and consumers*. Harpercollins.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61:65–170.

Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005-2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*. pages 57–60.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic, and morphological features. *Proceedings of COLING 2012* pages 1063–1080.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pages 690–696.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. volume 2, pages 174–180.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 328–339.

Frank Keller. 2001. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Ph.D. thesis, University of Edinburgh.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*. pages 562–568.

Gerasimos Lampouras and Andreas Vlachos. 2016. Imitation learning for language generation from unaligned data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. pages 1101–1112.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 228–231.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics pages 71–78.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics* 15(4):474–496.

André FT Martins, Fabio Kepler, and Jose Monteiro. 2017. Unbabel's participation in the WMT17 translation quality estimation shared task. In *Proceedings of the Second Conference on Machine Translation*. pages 569–574.

Chris Mellish and Robert Dale. 1998. Evaluation in the Context of Natural Language Generation. *Computer Speech & Language* 12(4):349–373. https://doi.org/10.1006/csla.1998.0106.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *CoRR* abs/1609.07843. http://arxiv.org/abs/1609.07843.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. pages 344–351.

Daniel Naber. 2003. A rule-based style and grammar checker .

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. *arXiv preprint arXiv:1610.02124* .

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017a. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 2241–2252. http://arxiv.org/abs/1707.06875.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017b. The e2e dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. pages 201–206.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 311–318.

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

Christian S Perone, Roberto Silveira, and Thomas S Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259* .

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. volume 1, pages 2227–2237.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training .

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics* 35(4):529–558.

Shikhar Sharma, Jing He, Kaheer Suleman, Hannes Schulz, and Philip Bachman. 2016. Natural language generation in dialogue using lexicalized and delexicalized data. *arXiv preprint arXiv:1606.03632* .

Anastasia Shimorina. 2018. Human vs automatic metrics: on the importance of correlation design. *arXiv preprint arXiv:1805.11474* .

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with QuEst++. *Proceedings of ACL-IJCNLP 2015 System Demonstrations* pages 115–120.

Lynne M Stallings, Maryellen C MacDonald, and Padraig G O'Seaghdha. 1998. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-np shift. *Journal of Memory and Language* 39(3):392–417.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pages 341–351.

David Temperley. 2008. Dependency-length minimization in natural and artificial languages. *Journal of Quantitative Linguistics* 15(3):256–282.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*. Association for Computational Linguistics, Stroudsburg, PA, USA, ConLL '00, pages 127–132. https://doi.org/10.3115/1117601.1117631.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. pages 163–173.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 4566–4575.

Joachim Wagner, Jennifer Foster, and Josef van Genabith. 2006. Detecting grammatical errors using probabilistic parsing. In *Workshop on Interfaces of Intelligent Computer-Assisted Language Learning*. pages 1–25. http://www.noe-kaleidoscope.org/group/idill/repository/iicall06.pdf.

Joachim Wagner, Jennifer Foster, and Josef van Genabith. 2009. Judging grammaticality: Experiments in sentence classification. *CALICO Journal* 26(3):474–490.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1711–1721. http://aclweb.org/anthology/D15-1199.