GLEU: Automatic Evaluation of Sentence-Level Fluency

Andrew Mutton* **Mark Dras*** **Stephen Wan*,†** **Robert Dale***

*Centre for Language Technology
Macquarie University
NSW 2109 Australia

†Information and Communication Technologies
CSIRO
NSW 2109 Australia

madras@ics.mq.edu.au

Abstract

In evaluating the output of language technology applications—MT, natural language generation, summarisation—automatic evaluation techniques generally conflate measurement of faithfulness to source content with fluency of the resulting text. In this paper we develop an automatic evaluation metric to estimate fluency alone, by examining the use of parser outputs as metrics, and show that they correlate with human judgements of generated text fluency. We then develop a machine learner based on these, and show that this performs better than the individual parser metrics, approaching a lower bound on human performance. We finally look at different language models for generating sentences, and show that while individual parser metrics can be 'fooled' depending on generation method, the machine learner provides a consistent estimator of fluency.

1 Introduction

Intrinsic evaluation of the output of many language technologies can be characterised as having at least two aspects: how well the generated text reflects the source data, whether it be text in another language for machine translation (MT), a natural language generation (NLG) input representation, a document to be summarised, and so on; and how well it conforms to normal human language usage. These two aspects are often made explicit in approaches to creating the text. For example, in statistical MT the translation model and the language model are treated separately, characterised as faithfulness and fluency respectively (as in the treatment in Jurafsky and Martin (2000)). Similarly, the ultrasummarisation model of Witbrock and Mittal (1999) consists of a content model, modelling the probability that a word in the source text will be in the summary, and a language model.

Evaluation methods can be said to fall into two categories: a comparison to gold reference, or an appeal to human judgements. Automatic evaluation methods carrying out a comparison to gold reference tend to conflate the two aspects of faithfulness and fluency in giving a goodness score for generated output. BLEU (Papineni et al., 2002) is a canonical example: in matching n-grams in a candidate translation text with those in a reference text, the metric measures faithfulness by counting the matches, and fluency by implicitly using the reference n-grams as a language model. Often we are interested in knowing the quality of the two aspects separately; many human judgement frameworks ask specifically for separate judgements on elements of the task that correspond to faithfulness and to fluency. In addition, the need for reference texts for an evaluation metric can be problematic, and intuitively seems unnecessary for characterising an aspect of text quality that is not related to its content source but to the use of language itself. It is a goal of this paper to provide an automatic evaluation method for fluency alone, without the use of a reference text.

One might consider using a metric based on language model probabilities for sentences: in eval-

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 344–351,
Prague, Czech Republic, June 2007.

uating a language model on (already existing) test data, a higher probability for a sentence (and lower perplexity over a whole test corpus) indicates better language modelling; perhaps a higher probability might indicate a better sentence. However, here we are looking at generated sentences, which have been generated using their own language model, rather than human-authored sentences already existing in a test corpus; and so it is not obvious what language model would be an objective assessment of sentence naturalness. In the case of evaluating a single system, using the language model that generated the sentence will only confirm that the sentence does fit the language model; in situations such as comparing two systems which each generate text using a different language model, it is not obvious that there is a principled way of deciding on a fair language model. Quite a different idea was suggested in Wan et al. (2005), of using the grammatical judgement of a parser to assess fluency, giving a measure independent of the language model used to generate the text. The idea is that, assuming the parser has been trained on an appropriate corpus, the poor performance of the parser on one sentence relative to another might be an indicator of some degree of ungrammaticality and possibly disfluency. In that work, however, correlation with human judgements was left uninvestigated.

The goal of this paper is to take this idea and develop it. In Section 2 we look at some related work on metrics, in particular for NLG. In Section 3, we verify whether parser outputs can be used as estimators of generated sentence fluency by correlating them with human judgements. In Section 4, we propose an SVM-based metric using parser outputs as features, and compare its correlation against human judgements with that of the individual parsers. In Section 5, we investigate the effects on the various metrics from different types of language model for the generated text. Then in Section 6 we conclude.

2 Related Work

In terms of human evaluation, there is no uniform view on what constitutes the notion of fluency, or its relationship to grammaticality or similar concepts. We mention a few examples here to illustrate the range of usage. In MT, the 2005 NIST MT Evaluation Plan uses guidelines[1] for judges to assess 'adequacy' and 'fluency' on 5 point scales, where they are asked to provide intuitive reactions rather than pondering their decisions; for fluency, the scale descriptions are fairly vague (5: flawless English; 4: good English; 3: non-native English; 2: disfluent English; 1: incomprehensible) and instructions are short, with some examples provided in appendices. Zajic et al. (2002) use similar scales for summarisation. By contrast, Pan and Shaw (2004), for their NLG system SEGUE tied the notion of fluency more tightly to grammaticality, giving two human evaluators three grade options: good, minor grammatical error, major grammatical/pragmatic error. As a further contrast, the analysis of Coch (1996) was very comprehensive and fine-grained, in a comparison of three text-production techniques: he used 14 human judges, each judging 60 letters (20 per generation system), and required them to assess the letters for correct spelling, good grammar, rhythm and flow, appropriateness of tone, and several other specific characteristics of good text.

In terms of automatic evaluation, we are not aware of any technique that measures only fluency or similar characteristics, ignoring content, apart from that of Wan et al. (2005). Even in NLG, where, given the variability of the input representations (and hence difficulty in verifying faithfulness), it might be expected that such measures would be available, the available metrics still conflate content and form. For example, the metrics proposed in Bangalore et al. (2000), such as Simple Accuracy and Generation Accuracy, measure changes with respect to a reference string based on the idea of string-edit distance. Similarly, BLEU has been used in NLG, for example by Langkilde-Geary (2002).

3 Parsers as Evaluators

There are three parts to verifying the usefulness of parsers as evaluators: choosing the parsers and the metrics derived from them; generating some texts for human and parser evaluation; and, the key part, getting human judgements on these texts and correlating them with parser metrics.

[1] http://projects.ldc.upenn.edu/TIDES/Translation/TranAssessSpec.pdf

3.1 The Parsers

In testing the idea of using parsers to judge fluency, we use three parsers, from which we derive four parser metrics, to investigate the general applicability of the idea. Those chosen were the Connexor parser,[2] the Collins parser (Collins, 1999), and the Link Grammar parser (Grinberg et al., 1995). Each produces output that can be taken as representing degree of ungrammaticality, although this output is quite different for each.

Connexor is a commercially available dependency parser that returns head–dependant relations as well as stemming information, part of speech, and so on. In the case of an ungrammatical sentence, Connexor returns tree fragments, where these fragments are defined by transitive head–dependant relations: for example, for the sentence *Everybody likes big cakes do* it returns fragments for *Everybody likes big cakes* and for *do*. We expect that the number of fragments should correlate inversely with the quality of a sentence. For a metric, we normalise this number by the largest number of fragments for a given data set. (Normalisation matters most for the machine learner in Section 4.)

The Collins parser is a statistical chart parser that aims to maximise the probability of a parse using dynamic programming. The parse tree produced is annotated with log probabilities, including one for the whole tree. In the case of ungrammatical sentences, the parser will assign a low probability to any parse, including the most likely one. We expect that the log probability (becoming more negative as the sentence is less likely) should correlate positively with the quality of a sentence. For a metric, we normalise this by the most negative value for a given data set.

Like Connexor, the Link Grammar parser returns information about word relationships, forming links, with the proviso that links cannot cross and that in a grammatical sentence all links are indirectly connected. For an ungrammatical sentence, the parser will delete words until it can produce a parse; the number it deletes is called the 'null count'. We expect that this should correlate inversely with sentence quality. For a metric, we normalise this by the sentence length. In addition, the parser produces another variable possibly of interest. In generating a parse, the parser produces many candidates and rules some out by a posteriori constraints on valid parses. In its output the parser returns the number of invalid parses. For an ungrammatical sentence, this number may be higher; however, there may also be more parses. For a metric, we normalise this by the total number of parses found for the sentence. There is no strong intuition about the direction of correlation here, but we investigate it in any case.

3.2 Text Generation Method

To test whether these parsers are able to discriminate sentence-length texts of varying degrees of fluency, we need first to generate texts that we expect will be discriminable in fluency quality ranging from good to very poor. Below we describe our method for generating text, and then our preliminary check on the discriminability of the data before giving them to human judges.

Our approach to generating 'sentences' of a fixed length is to take word sequences of different lengths from a corpus and glue them together probabilistically: the intuition is that a few longer sequences glued together will be more fluent than many shorter sequences. More precisely, to generate a sentence of length n, we take sequences of length l (such that l divides n), with sequence i of the form $w_{i,1} \ldots w_{i,l}$, where $w_{i,\_}$ is a word or punctuation mark. We start by selecting sequence 1, first by randomly choosing its first word according to the unigram probability $P(w_{1,1})$, and then the sequence uniformly randomly over all sequences of length l starting with $w_{1,1}$; we select subsequent sequences j ($2 \leq j \leq n/l$) randomly according to the bigram probability $P(w_{j,1} \mid w_{j-1,l})$. Taking as our corpus the Reuters corpus,[3] for length $n = 24$, we generate sentences for sequence sizes $l = 1, 2, 4, 8, 24$ as in Figure 1. So, for instance, the sequence-size 8 example was constructed by stringing together the three consecutive sequences of length 8 (*There* ... *to*; *be* ... *have*; *to* ....) taken from the corpus.

These examples, and others generated, appear to be of variable quality in accordance with our intuition. However, to confirm this prior to testing them

[2] http://www.connexor.com

[3] http://trec.nist.gov/data/reuters/reuters.html

Extracted (Sequence-size 24)
Ginebra face Formula Shell in a sudden-death playoff on Sunday to decide who will face Alaska in a best-of-seven series for the title.
Sequence-size 8
There is some thinking in the government to be nearly as dramatic as some people have to be slaughtered to eradicate the epidemic.
Sequence-size 4
Most of Banharn's move comes after it can still be averted the crash if it should again become a police statement said.
Sequence-size 2
Massey said in line with losses, Nordbanken is well-placed to benefit abuse was loaded with Czech prime minister Andris Shkele, said.
Sequence-size 1
The war we're here in a spokesman Jeff Sluman 86 percent jump that Spain to what was booked, express also said.

Figure 1: Sample sentences from the first trial

Description	Correlation
Small	0.10 to 0.29
Medium	0.30 to 0.49
Large	0.50 to 1.00

Table 1: Correlation coefficient interpretation

out for discriminability in a human trial, we wanted see whether they are discriminable by some method other than our own judgement. We used the parsers described in Section 3.1, in the hope of finding a non-zero correlation between the parser outputs and the sequence lengths.

Regarding the interpretation of the absolute value of (Pearson's) correlation coefficients, both here and in the rest of the paper, we adopt Cohen's scale (Cohen, 1988) for use in human judgements, given in Table 1; we use this as most of this work is to do with human judgements of fluency. For data, we generated 1000 sentences of length 24 for each sequence length $l = 1, 2, 3, 4, 6, 8, 24$, giving 7000 sentences in total. The correlations with the four parser outputs are as in Table 2, with the medium correlations for Collins and Link Grammar (nulled tokens) indicating that the sentences are indeed discriminable to some extent, and hence the approach is likely to be useful for generating sentences for human trials.

3.3 Human Judgements

The next step is then to obtain a set of human judgements for this data. Human judges can only be expected to judge a reasonably sized amount of data,

Metric	Corr.
Collins Parser	0.3101
Connexor	-0.2332
Link Grammar Nulled Tokens	-0.3204
Link Grammar Invalid Parses	0.1776
GLEU	0.4144

Table 2: Parser vs sequence size for original data set

so we first reduced the set of sequence sizes to be judged. To do this we determined for the 7000 generated sentences the scores according to the (arbitrarily chosen) Collins parser, and calculated the means for each sequence size and the 95% confidence intervals around these means. We then chose a subset of sequence sizes such that the confidence intervals did not overlap: $1, 2, 4, 8, 24$; the idea was that this would be likely to give maximally discriminable sentences. For each of these sequences sizes, we chose randomly 10 sentences from the initial set, giving a set for human judgement of size 50.

The judges consisted of twenty volunteers, all native English speakers without explicit linguistic training. We gave them general guidelines about what constituted fluency, mentioning that they should consider grammaticality but deliberately not giving detailed instructions on the manner for doing this, as we were interested in the level of agreement of intuitive understanding of fluency. We instructed them also that they should evaluate the sentence without considering its content, using *Colourless green ideas sleep furiously* as an example of a nonsensical but perfectly fluent sentence. The judges were then presented with the 50 sentences in random order, and asked to score the sentences according to their own scale, as in magnitude estimation (Bard et al., 1996); these scores were then normalised in the range [0,1].

Some judges noted that the task was difficult because of its subjectivity. Notwithstanding this subjectivity and variation in their approach to the task, the pairwise correlations between judges were high, as indicated by the maximum, minimum and mean values in Table 3, indicating that our assumption that humans had an intuitive notion of fluency and needed only minimal instruction was justified. Looking at mean scores for each sequence size, judges generally also ranked sentences by sequence size; see Figure 2. Comparing human judgement

Statistic	Corr.
Maximum correlation	0.8749
Minimum correlation	0.4710
Mean correlation	0.7040
Standard deviation	0.0813

Table 3: Data on correlation between humans

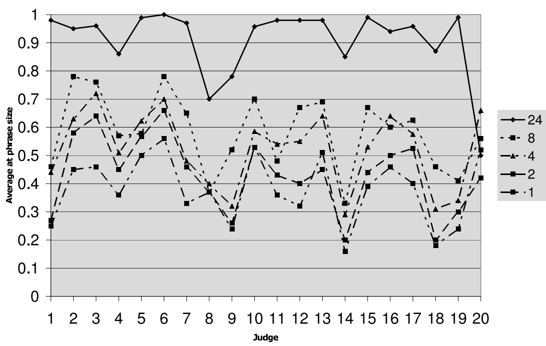

Figure 2: Mean scores for human judges

correlations against sequence size with the same correlations for the parser metrics (as for Table 2, but on the human trial data) gives Table 4, indicating that humans can also discriminate the different generated sentence types, in fact (not surprisingly) better than the automatic metrics.

Now, having both human judgement scores of some reliability for sentences, and scoring metrics from three parsers, we give correlations in Table 5. Given Cohen's interpretation, the Collins and Link Grammar (nulled tokens) metrics show moderate correlation, the Connexor metric almost so; the Link Grammar (invalid parses) metric correlation is by far the weakest. The consistency and magnitude of the first three parser metrics, however, lends support to the idea of Wan et al. (2005) to use something like these as indicators of generated sentence fluency. The aim of the next section is to build a better predictor than the individual parser metrics alone.

Metric	Corr.
Humans	0.6529
Collins Parser	0.4057
Connexor	-0.3804
Link Grammar Nulled Tokens	-0.3310
Link Grammar Invalid Parses	0.1619
GLEU	0.4606

Table 4: Correlation with sequence size for human trial data set

Metric	Corr.
Collins Parser	0.3057
Connexor	-0.3445
Link-Grammar Nulled Tokens	-0.2939
Link Grammar Invalid Parses	0.1854
GLEU	0.4014

Table 5: Correlation between metrics and human evaluators

4 An SVM-Based Metric

In MT, one problem with most metrics like BLEU is that they are intended to apply only to document-length texts, and any application to individual sentences is inaccurate and correlates poorly with human judgements. A neat solution to poor sentence-level evaluation proposed by Kulesza and Shieber (2004) is to use a Support Vector Machine, using features such as word error rate, to estimate sentence-level translation quality. The two main insights in applying SVMs here are, first, noting that human translations are generally good and machine translations poor, that binary training data can be created by taking the human translations as positive training instances and machine translations as negative ones; and second, that a non-binary metric of translation goodness can be derived by the distance from a test instance to the support vectors. In an empirical evaluation, Kulesza and Shieber found that their SVM gave a correlation of 0.37, which was an improvement of around half the gap between the BLEU correlations with the human judgements (0.25) and the lowest pairwise human inter-judge correlation (0.46) (Turian et al., 2003).

We take a similar approach here, using as features the four parser metrics described in Section 3. We trained an SVM,[4] taking as positive training data the 1000 instances of sentences of sequence length 24 (i.e. sentences extracted from the corpus) and as negative training data the 1000 sentences of sequence length 1. We call this learner GLEU.[5]

As a check on the ability of the GLEU SVM to distinguish these 'positive' sentences from 'negative' ones, we evaluated its classification accuracy on a (new) test set of size 300, split evenly between sentences of sequence length 24 and sequence length 1.

[4]We used the package SVM-light (Joachims, 1999).

[5]For GrammaticaLity Evaluation Utility.

This gave 81%, against a random baseline of 50%, indicating that the SVM can classify satisfactorily.

We now move from looking at classification accuracy to the main purpose of the SVM, using distance from support vector as a metric. Results are given for correlation of GLEU against sequence sizes for all data (Table 2) and for the human trial data set (Table 4); and also for correlation of GLEU against the human judges' scores (Table 5). This last indicates that GLEU correlates better with human judgements than any of the parsers individually, and is well within the 'moderate' range for correlation interpretation. In particular, for the GLEU–human correlation, the score of 0.4014 is approaching the minimum pairwise human correlation of 0.4710.

5 Different Text Generation Methods

The method used to generate text in Section 3.2 is a variation of the standard n-gram language model. A question that arises is: Are any of the metrics defined above strongly influenced by the type of language model used to generate the text? It may be the case, for example, that a parser implementation uses its own language model that predisposes it to favour a similar model in the text generation process. This is a phenomenon seen in MT, where BLEU seems to favour text that has been produced using a similar statistical n-gram language model over other symbolic models (Callison-Burch et al., 2006).

Our previous approach used only sequences of words concatenated together. To define some new methods for generating text, we introduced varying amounts of structure into the generation process.

5.1 Structural Generation Methods

PoStag In the first of these, we constructed a rough approximation of typical sentence grammar structure by taking bigrams over part-of-speech tags.[6] Then, given a string of PoS tags of length n, $t_1 \dots t_n$, we start by assigning the probabilities for the word in position 1, w_1, according to the conditional probability $P(w_1 \,|\, t_1)$. Then, for position j ($2 \leq j \leq n$), we assign to candidate words the value $P(w_j \,|\, t_j) \times P(w_j \,|\, w_{j-1})$ to score word sequences. So, for example, we might generate the PoS tag template **Det NN Adj Adv**, take all the words corresponding to each of these parts of speech, and combine bigram word sequence probability with the conditional probability of words with respect to these parts of speech. We then use a Viterbi-style algorithm to find the most likely word sequence.

In this model we violate the Markov assumption of independence in much the same way as Witbrock and Mittal (1999) in their combination of content and language model probabilities, by backtracking at every state in order to discourage repeated words and avoid loops.

Supertag This is a variant of the approach above, but using supertags (Bangalore and Joshi, 1999) instead of PoS tags. The idea is that the supertags might give a more fine-grained definition of structure, using partial trees rather than parts of speech.

CFG We extracted a CFG from the ∼10% of the Penn Treebank found in the NLTK-lite corpora.[7] This CFG was then augmented with productions derived from the PoS-tagged data used above. We then generated a template of length n pre-terminal categories using this CFG. To avoid loops we biased the selection towards terminals over non-terminals.

5.2 Human Judgements

We generated sentences according to a mix of the initial method of Section 3.2, for calibration, and the new methods above. We again used a sentence length of 24, and sequence lengths for the initial method of $l = 1, 8, 24$. A sample of sentences generated for each of these six types is in Figure 3.

For our data, we generated 1000 sentences per generation method, giving a corpus of 6000 sentences. For the human judgements we also again took 10 sentences per generation method, giving 60 sentences in total. The same judges were given the same instructions as previously.

Before correlating the human judges' scores and the parser outputs, it is interesting to look at how each parser treats the sentence generation methods, and how this compares with human ratings (Table 6). In particular, note that the Collins parser rates the PoStag- and Supertag-generated sentences more

[6] We used the supertagger of Bangalore and Joshi (1999).

[7] `http://nltk.sourceforge.net`

Extracted (Sequence-size 24)
After a near three-hour meeting and last-minute talks with President Lennart Meri, the Reform Party council voted overwhelmingly to leave the government.
Sequence-size 8
If Denmark is closely linked to the Euro Disney reported a net profit of 85 million note: the figures were rounded off.
Sequence-size 1
Israelis there would seek approval for all-party peace now complain that this year, which shows demand following year and 56 billion pounds.
POS-tag, Viterbi-mapped
He said earlier the 9 years and holding company's government, including 69.62 points as a number of last year but market.
Supertag, Viterbi-mapped
That 97 saying he said in its shares of the market 74.53 percent, adding to allow foreign exchange: I think people.
Context-Free Grammar
The production moderated Chernomyrdin which leveled government back near own 52 over every a current at from the said by later the other.

Figure 3: Sample sentences from the second trial

sent. type	s-24	s-8	s-1	PoS	sup.	CFG
Collins	**0.52**	0.48	0.41	**0.60**	**0.57**	0.36
Connexor	0.12	0.16	0.24	0.26	0.25	0.43
LG (null)	0.02	0.06	0.10	0.09	0.11	0.18
LG (invalid)	0.78	0.67	0.56	0.62	0.66	0.53
GLEU	1.07	0.32	-0.96	0.28	-0.06	-2.48
Human	0.93	0.67	0.44	0.39	0.44	0.31

Table 6: Mean normalised scores per sentence type

highly even than real sentences (in bold). These are the two methods that use the Viterbi-style algorithm, suggesting that this probability maximisation has fooled the Collins parser. The pairwise correlation between judges was around the same on average as in Section 3.3, but with wider variation (Table 7).

The main results, determining the correlation of the various parser metrics plus GLEU against the new data, are in Table 8. This confirms the very variable performance of the Collins parser, which has dropped significantly. GLEU performs quite consistently here, this time a little behind the Link Grammar (nulled tokens) result, but still with a better correlation with human judgement than at least two

Statistic	Corr.
Maximum correlation	0.9048
Minimum correlation	0.3318
Mean correlation	0.7250
Standard deviation	0.0980

Table 7: Data on correlation between humans

Metric	Corr.
Collins Parser	0.1898
Connexor	-0.3632
Link-Grammar Nulled Tokens	-0.4803
Link Grammar Invalid Parses	0.1774
GLEU	0.4738

Table 8: Correlation between parsers and human evaluators on new human trial data

Metric	Corr.
Collins Parser	0.2313
Connexor	-0.2042
Link-Grammar Nulled Tokens	-0.1289
Link Grammar Invalid Parses	-0.0084
GLEU	0.4312

Table 9: Correlation between parsers and human evaluators on all human trial data

judges with each other. (Note also that the GLEU SVM was not retrained on the new sentence types.)

Looking at all the data together, however, is where GLEU particularly displays its consistency. Aggregating the old human trial data (Section 3.3) and the new data, and determining correlations against the metrics, we get the data in Table 9. Again the SVM's performance is consistent, but is now almost twice as high as its nearest alternative, Collins.

5.3 Discussion

In general, there is at least one parser that correlates quite well with the human judges for each sentence type. With well-structured sentences, the probabilistic Collins parser performs best; on sentences that are generated by a poor probabilistic model leading to poor structure, Link Grammar (nulled tokens) performs best. This supports the use of a machine learner taking as features outputs from several parser types; empirically this is confirmed by the large advantage GLEU has on overall data (Table 9).

The generated text itself from the Viterbi-based generators as implemented here is quite disappointing, given an expectation that introducing structure would make sentences more natural and hence lead to a range of sentence qualities. In hindsight, this is not so surprising; in generating the structure template, only sequences (over tags) of size 1 were used, which is perhaps why the human judges deemed them fairly close to sentences generated by the original method using sequence size 1, the poorest of that initial data set.

6 Conclusion

In this paper we have investigated a new approach to evaluating the fluency of individual generated sentences. The notion of what constitutes fluency is an imprecise one, but trials with human judges have shown that even if it cannot be exactly defined, or even articulated by the judges, there is a high level of agreement about what is fluent and what is not. Given this data, metrics derived from parser outputs have been found useful for measuring fluency, correlating up to moderately well with these human judgements. A better approach is to combine these in a machine learner, as in our SVM GLEU, which outperforms individual parser metrics. Interestingly, we have found that the parser metrics can be fooled by the method of sentence generation; GLEU, however, gives a consistent estimate of fluency regardless of generation type; and, across all types of generated sentences examined in this paper, is superior to individual parser metrics by a large margin.

This all suggests that the approach has promise, but it needs to be developed further for pratical use. The SVM presented in this paper has only four features; more features, and in particular a wider range of parsers, should raise correlations. In terms of the data, we looked only at sentences generated with several parameters fixed, such as sentence length, due to our limited pool of judges. In future we would like to examine the space of sentence types more fully. In particular, we will look at predicting the fluency of near-human quality sentences. More generally, we would like to look also at how the approach of this paper would relate to a perplexity-based metric; how it compares against BLEU or similar measures as a predictor of fluency in a context where reference sentences are available; and whether GLEU might be useful in applications such as reranking of candidate sentences in MT.

Acknowledgements

We thank Ben Hutchinson and Mirella Lapata for discussions, and Srinivas Bangalore for the TAG supertagger. The second author acknowledges the support of ARC Discovery Grant DP0558852.

References

Srinivas Bangalore and Aravind Joshi. 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265.

Srinivas Bangalore, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the First International Natural Language Generation Conference (INLG2000)*, Mitzpe Ramon, Israel.

E. Bard, D. Robertson, and A. Sorace. 1996. Magnitude estimation and linguistic acceptability. *Language*, 72(1):32–68.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In *Proceedings of EACL*, pages 249–256.

José Coch. 1996. Evaluating and comparing three text-production strategies. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 249–254.

J. Cohen. 1988. *Statistical power analysis for the behavioral sciences*. Erlbaum, Hillsdale, NJ, US.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Dennis Grinberg, John Lafferty, and Daniel Sleator. 1995. A robus parsing algorithm for link grammars. In *Proceedings of the Fourth International Workshop on Parsing Technologies*.

Thorsten Joachims. 1999. *Making Large-Scale SVM Learning Practical*. MIT Press.

Daniel Jurafsky and James Martin. 2000. *Speech and Language Processing: An Introduction to Natural Languge Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall.

Alex Kulesza and Stuart Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, US.

Irene Langkilde-Geary. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the International Natural Language Generation Conference (INLG) 2002*, pages 17–24.

Shimei Pan and James Shaw. 2004. Segue: A hybrid case-based surface natural language generator. In *Proceedings of the International Conference on Natural Language Generation (INLG) 2004*, pages 130–140.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176, IBM.

Joseph Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of Machine Translation and its evaluation. In *Proceedings of MT Summit IX*, pages 23–28.

Stephen Wan, Robert Dale, Mark Dras, and Cécile Paris. 2005. Searching for grammaticality: Propagating dependencies in the Viterbi algorithm. In *Proceedings of the 10th European Natural Language Processing Wworkshop*, Aberdeen, UK.

Michael Witbrock and Vibhu Mittal. 1999. Ultra-summarization: A statistical approach to generating highly condensed non-executive summaries. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*.

David Zajic, Bonnie Dorr, and Richard Schwartz. 2002. Automatic headline generation for newspaper stories. In *Proceedings of the ACL-2002 Workshop on Text Summarization (DUC2002)*, pages 78–85.