Evaluating Evaluation Methods for Generation in the Presence of Variation

Amanda Stent, Matthew Marge, and Mohit Singhai

Stony Brook University, Stony Brook, NY 11794, USA,
stent@cs.sunysb.edu, mmarge@ic.sunysb.edu, mohit@cs.sunysb.edu

Abstract. Recent years have seen increasing interest in automatic metrics for the evaluation of generation systems. When a system can generate syntactic variation, automatic evaluation becomes more difficult. In this paper, we compare the performance of several automatic evaluation metrics using a corpus of automatically generated paraphrases. We show that these evaluation metrics can at least partially measure adequacy (similarity in meaning), but are not good measures of fluency (syntactic correctness). We make several proposals for improving the evaluation of generation systems that produce variation.

1 Introduction

The task of surface realization is to select, inflect and order words to communicate the input meaning as completely, clearly and fluently as possible in context. Traditional grammar-based surface realizers, (*e.g.* [1]) focus on the production of at least one high quality output sentence for each input semantic form. By contrast, two-stage surface realizers (e.g. [2, 3]) produce many possible sentences for each input semantic form, but select only one for output. Comparatively little research has been performed on rule-based approaches to the generation of variation (but see [4, 5]). However, recently there has been increasing interest on corpus-based approaches to the generation of paraphrases, or text-to-text generation (e.g. [6–10]).

Variation in surface realization takes two basic forms: *word choice variation*, and *word order variation*. Example 1 shows both types of variation. Word order variation may entail word choice variation, as in example (1b).

Example 1
(a) I bought tickets for the show on Tuesday.
(b) It was the show on Tuesday for which I bought tickets.
(c) I got tickets for the show on Tuesday.
(d) I bought tickets for the Tuesday show.
(e) On Tuesday I bought tickets for the show.
(f) For the show on Tuesday tickets I bought.

Variation is widely used by humans both in text and dialog. However, not all variations are meaning-preserving. A variation may add meaning possibilities that were not there before, remove meaning possibilities (compare example (1a) with (1d) and

(1e)), or otherwise change the meaning of part of a sentence. As example (1f) shows, a variation may also be unclear or syntactically incorrect.

In this paper, we will say that a *valid variation* of a sentence must meet three criteria: *adequacy*, or meaning equivalence; *fluency*, or syntactic correctness; and *readability*, or efficacy in context[1]. A sentence that is ambiguous, that does not express all the input meaning, or that communicates meaning not contained in the input, is not adequate. Even if a sentence is adequate, if it is not syntactically correct or idiomatic it is not fluent. Sentences that are both adequate and fluent may still not be adequate or fluent in a particular context; they are not readable in that context. Pronouns and discourse cues are two constructs that may affect the readability of a sentence.

It may seem odd to separate readability from adequacy and fluency, as context can affect both adequacy and fluency. For example, context can be used for disambiguation. However, very few surface realizers and no automatic evaluation metrics take context into account. The influence of discourse context in surface realization remains an important but poorly understood topic.

Most automatic evaluation metrics for generation and machine translation do not directly evaluate adequacy or fluency; rather, they indirectly evaluate these criteria by comparing the generated, or *candidate*, sentence to one or more human-created *reference* sentences [11–14]. Where a metric permits the comparison of a candidate sentence to multiple reference sentences [11–13], only one reference sentence is typically used in evaluation of generation quality ([15, 16], c.f. [14]). Tree-based metrics incorporating the notion of constituency, *e.g.* [14], are not widely used because they require correct parse trees for reference and candidate sentences. No existing automatic evaluation metric evaluates readability.

An interesting question is the extent to which automatic evaluation metrics for text generation systems can be used to evaluate the output quality of generation systems that produce variation. A good automatic evaluation metric could be useful not only for evaluation and comparison of generation systems, but also for distinguishing valid from invalid variations in the output of a two-stage surface realizer. This would permit greater flexibility and efficiency in two-stage surface realization.

Because existing automatic evaluation metrics for generation evaluate by comparison to one or more reference sentences rather than evaluating adequacy and fluency directly, they will punish both word choice and word order variation. However, it may be that they can still distinguish to some extent between valid and invalid sentences, *i.e.* that the noise introduced by variation is not sufficient to drown out the signal of validity. The question we address in this paper is whether existing automatic evaluation metrics for text generation can accurately evaluate the adequacy and fluency of generated sentences when variation is permitted. Sections 2 and 3 describe the data and metrics we used. Section 4 describes an experiment we conducted comparing the performance of these metrics to human judgments of adequacy and fluency. Section 5 discusses the implications of our results and our proposals for evaluation of generation systems that permit variation. We conclude with some ideas for future work.

[1] These are very similar to criteria used in machine translation evaluations, *e.g.* [11].

2 Data

The data we used for this study consists of a set of 118 automatically-generated paraphrase sentences made available by Barzilay and Lee[2]. Barzilay and Lee employ a corpus-based approach to paraphrase generation [6]. Sentences in a corpus are grouped by similarity, and then the *multiple sequence alignment* of each group of sentences is computed. The multiple sequence alignment of a group of sentences is a word lattice capturing places where the sentences are the same and places where they differ; it is a compact representation of possible variations of a sentence. A paraphrase is generated for a new input sentence by aligning the input sentence with one of the word lattices and then choosing an alternative path through that lattice.

The data we used includes sentences produced by Barzilay and Lee's baseline system (50%) and sentences produced by Barzilay and Lee's multiple sequence alignment based (MSA) system (50%). The baseline system simply replaces words in a sentence with one of their WordNet synonyms, at a rate proportional to the word replacement rate of the MSA system for that sentence. Therefore, the baseline system includes word choice variation only (example 2), whereas the MSA system includes both word choice and word order variation (example 3).

Example 2
(a) Another person was also seriously wounded in the attack.
(b) Another individual was also seriously wounded in the attack.

Example 3
(a) A suicide bomber blew himself up at a bus stop east of Tel Aviv on Thursday, killing himself and wounding five bystanders, one of them seriously, police and paramedics said.
(b) A suicide bomber killed himself and wounded five, when he blew himself up at a bus stop east of Tel Aviv on Thursday.

The variations produced using multiple sequence alignment are of very high quality and are typically highly fluent. However, because there is no explicit representation of the meaning of the input sentence, words chosen may occasionally carry connotations not carried by the words they replace, and sometimes words are included or removed that alter the meaning of the sentence (*e.g.* example 3).

3 Evaluation Metrics

We used five evaluation metrics for this study: NIST simple string accuracy (SSA) [14], the BLEU and NIST n-gram co-occurrence metrics [12, 11], Melamed's F measure [13], and latent semantic analysis (LSA) [17]. Only SSA and BLEU have previously been used to evaluate the output of generation systems; SSA, BLEU, NIST and the F measure are designed for the evaluation of machine translation output. As Table 1 shows,

[2] http://www.cs.cornell.edu/Info/Projects/NLP/statpar.html

all these metrics evaluate the fluency and adequacy of generated candidate sentences indirectly by comparison with one or more reference sentences. Table 1 also shows how one might use these metrics to evaluate readability, although we are not aware of any research that uses this approach.

Metric	**SSA**	**NIST n-gram, BLEU**	**F measure**	**LSA**
Means of measuring fluency	Comparison to reference sentence	Comparison to reference sentences – matching n-grams	Comparison to reference sentences – longest matching substrings	None
Means of measuring adequacy	Comparison to reference sentence	Comparison to reference sentences	Comparison to reference sentences	Comparison using word co-occurrence frequencies learned from corpus
Means of measuring readability	Comparison to reference sentence(s) from same context*	Comparison to reference sentences from same context*	Comparison to reference sentences from same context*	None
Punishes length differences?	Yes (punishes deletions, insertions)	Yes (weights)	Yes (weights)	Not explicitly

Table 1. Evaluation metrics

Simple String Accuracy The NIST simple string accuracy (SSA) metric scores a candidate sentence by tallying the number of substitutions, insertions, and deletions necessary to convert the reference sentence to the candidate sentence and dividing by the length of the candidate sentence. SSA has been used to evaluate the output of SURGE [16] and FERGUS [14].

BLEU IBM's BLEU metric, designed for evaluating machine translation quality, scores candidate sentences by counting the number of n-gram matches between candidate and reference sentences. It also punishes differences in length between candidate and reference sentences. The BLEU evaluation metric has been shown to correlate highly with human judgments [12]. The BLEU metric has been used to evaluate the output of HALogen [15].

NIST The NIST n-gram based evaluation metric, also designed for evaluating machine translation quality, differs from the BLEU metric in three ways. First, The arithmetic mean of co-occurrences is used instead of the geometric mean. Second, n-grams that occur less frequently are weighted more highly than those that occur more frequently. Third, there is a slightly different length penalty. These differences have been shown to lead to a higher correlation with human judgments than BLEU has [11]. Unlike the other metrics, NIST n-gram scores are not in the range $[0, 1]$. We include it primarily for comparison with BLEU.

F Measure This metric was developed by Melamed et. al. for evaluating machine translation quality [13]. It is designed to eliminate the "double counting" done by n-gram based metrics such as the NIST and BLEU n-gram based metrics (which penalize the same word insertion, deletion or movement as it occurs in a unigram, a bigram, etc.). It uses two scores, precision and recall, computed separately for each candidate sentence. Both precision and recall are defined in terms of the maximum match size, which is the weighted sum of the lengths of the longest matching text blocks between candidate and reference sentences. Precision is the maximum match size divided by the length of the candidate sentence; recall is the maximum match size divided by the length of the reference sentence. The maximum match size can be adjusted to weight longer matches more or less heavily by using a different exponent; for this data, 1 was the best exponent. Studies by Melamed et. al. show a high correlation between this metric and human judgments of translation quality [13]. This metric punishes variation in sentence length less than BLEU and NIST, so we hypothesized that it would be more closely correlated with human judgments for variation generation.

Latent Semantic Analysis Latent semantic analysis (LSA) computes the semantic similarity of two texts by measuring the semantic similarities of the words they contain [17]. Semantic similarity is computed by means of word co-occurrence counts obtained from a large corpus. LSA differs from the other metrics we used in two ways. First, it treats each sentence as a bag of words (compares sentences without regard to word order). Second, it uses word co-occurrence statistics learned from a large corpus to compute the semantic similarity of words. Therefore, we hypothesized that LSA would be good at evaluating adequacy in the presence of variation, although obviously it cannot serve as a measure of fluency.

4 Procedure

As Table 1 shows, most automatic evaluation metrics compare generated sentences to one or more reference sentences. This means that automatic evaluation metrics will tend to punish word choice and word order variation. The questions addressed in this experiment are: a) Are automatic evaluation metrics sufficiently robust to variation to distinguish between sentences that are valid (adequate and fluent) and those that are not?; and b) What is the relative impact of word choice and word order variation on the performance of these metrics? Our procedure was to compare human judgments of adequacy and fluency to the scores of the five selected evaluation metrics for the two sets of paraphrases provided by Barzilay and Lee.

We had three human judges evaluate the paraphrase pairs provided by Barzilay and Lee. In the following discussion, the reference sentence is the original (human-created) sentence, and the candidate sentence is an output from one of the two systems used by Barzilay and Lee.

Each judge answered two questions for each reference/candidate sentence pair, one pertaining to adequacy and one to fluency. For each sentence pair, the reference sentence is sentence A and the candidate sentence is sentence B. In our evaluation judges did not see the sentences in a larger discourse context, since the evaluation metrics do

	BLEU	NIST	SSA	F	LSA	Adequacy	Fluency
BLEU	1.00						
NIST	0.910	1.00					
SSA	0.894	0.863	1.00				
F	0.927	0.900	0.955	1.00			
LSA	0.725	0.727	0.742	0.795	1.00		
Adequacy	0.388	0.421	0.412	0.457	0.467	1.00	
Fluency	-0.492	-0.563	-0.400	-0.412	-0.290	*-0.032*	1.00
Length candidate	0.540	0.722	0.426	0.467	0.421	*0.169*	-0.374

Table 2. Correlation between human judgments of meaning preservation and syntactic accuracy and automatic evaluation metrics

not consider discourse context, so there is no evaluation of readability. The questions the judges were asked are:

1. How much of the meaning expressed in Sentence A is also expressed in Sentence B?
__All __Most __Half __Some __None
2. How do you judge the fluency of Sentence B? It is
__Flawless __Good __Adequate __Poor __Incomprehensible

The paraphrases were rated very highly in general. In the experiment reported below, the judges' ratings are averaged and normalized to the range [0,1]. The mean rating for adequacy was 4 (st. dev. 0.66, min. 2, max. 5), and for fluency was 4.13 (st. dev. 0.72, min. 2.33, max. 5).

All paraphrases were also evaluated using the five automatic evaluation metrics described in the previous section. We then computed the correlation between the human evaluations of adequacy and fluency and the scores for each evaluation metric. We used the Spearman rank coefficient of correlation, which is a measure of the strength of the linear relationship between two variables. We used Spearman rather than the Pearson coefficient because this data is not normally distributed.

5 Results

Our comparison of these evaluation metrics is shown in Table 2. All correlations are significant at $p < .01$ unless italicized. Correlations greater than 0.67 indicate strong relationships, while correlations between 0.34 and 0.66 indicate some relationship.

As one would expect, the automatic evaluation metrics are highly positively correlated with each other. Also as one would expect, the automatic evaluation metrics are positively correlated with the length of the candidate sentence.

There is no significant correlation between human judgments of adequacy and human judgments of fluency, indicating that the judges considered these two dimensions separately. Because the judges could always see both the candidate and the reference sentences, they may have tended to make slightly higher judgments of adequacy than

they would have otherwise. On the other hand, the paraphrases are of very high quality in general, even when meaning is not completely preserved. It should be noted that the median difference in sentence length between source and target sentences was 2 words; i.e. generated sentences were not usually summaries of the input sentences, so typically most of the meaning was preserved. There were cases where information was added, but it was typically attribution information (*e.g. police said*).

Adequacy There are positive, but not strong, correlations between the scores of the automatic evaluation metrics and human judgments of adequacy. We conclude that *these automatic evaluation metrics are adequate, but not good, evaluators of adequacy.*

Fluency There are negative correlations between the scores of the automatic evaluation metrics and human judgments of fluency. This is weakest in the case of LSA, which does not consider word order. We conclude that (at least in the presence of variation) *these automatic evaluation metrics are poor evaluators of fluency.*

System	BLEU	NIST	SSA	F	LSA	Adequacy	Fluency
Baseline	.753	4.156	.864	.888	.954	*.833*	.756
MSA	.290	1.945	.423	.530	.845	*.770*	.897

Table 3. Baseline vs. Multiple Sequence Alignment

Impact of word order variation Recall that Barzilay and Lee's baseline system performs word choice variation only, while their MSA system performs both word choice and word order variation. Furthermore, the frequency of word choice variation was held constant across both systems. Therefore, this data set is useful for evaluating the relative impact of word order variation on automatic evaluation scores.

The means of the scores for each system are shown in Table 3. A paired t-test showed that these differences are significant at $p < .01$ for all except human adequacy judgments. The automatic evaluation metrics all scored the baseline system *higher* than the MSA system. In contrast, the human judges rated the fluency of the MSA system output higher than that of the baseline system. Mostly, this is because the MSA system can make decisions about word choice variation based on the context in which the word appears while the baseline system cannot; however, sometimes the MSA system produced paraphrases that were clearly more readable than the input sentence. The human judges rated the output of both systems highly for adequacy. We conclude that, because these evaluation metrics punish word order (and word choice) variation in ways that do not distinguish between valid and invalid variations, *these automatic evaluation metrics are not adequate for the task of evaluating variation generation*.

6 Discussion

The results of the experiment in the previous section demonstrate that existing automatic evaluation metrics are inadequate for evaluating the output of generation systems

that produce variation. In this section, we discuss four proposals for improving the automatic evaluation of generation systems that produce variation.

Proposal 1: Multiple reference sentences Several of the metrics used in this paper (*e.g.* [11–13]) permit multiple reference sentences. Where it is possible to find multiple reference sentences covering the range of possible variants on a sentence, these metrics might prove more closely correlated with human judgments of adequacy and fluency. We therefore recommend that *automatic evaluations of the quality of surface realizers should be conducted using multiple reference sentences*.

This recommendation comes with two caveats. First, it can be time-consuming to find multiple reference sentences for each sentence in a test set for a particular domain, and an out-of-domain test set may not provide an honest accounting of the quality of the surface realizer output. Second, even if multiple reference sentences are provided, two problems remain: a) it is highly likely that some valid variations will not be included, and b) it is possible for two variations of parts of a sentence to be fluent and adequate separately, but not in combination, as example 4 shows:

Example 4
(a) She killed her with a gunshot to the head.
(b) She shot her in the head.
→ (c) She shot her to the head with a gunshot.

Proposal 2: Shallow models of constituency As Callaway points out in [16], most automatic evaluation metrics for generation do not contain models of syntactic constituency. This is a serious drawback when it comes to evaluating generation systems that permit variation. In particular, the lack of a model of constituency means that automatic evaluation metrics cannot distinguish between valid movement such as that in example 1b and invalid movement (for example, *I bought tickets on Tuesday the show for.*).

There are three possible solutions to this problem: use a parser to evaluate the fluency of generated sentences, use a grammar checker to evaluate fluency, or use tree-based evaluation metrics. Unfortunately, since parsers are descriptive rather than prescriptive models of language, they are not suitable for evaluation purposes. We tried parsing a set of fluent and disfluent permutations of the words in example (1a) using the Collins and Charniak parsers, and obtained parses for all of them. Furthermore, the probabilities assigned to some of the very disfluent parses were higher than those assigned to some of the less disfluent ones.

Similarly, grammar checkers do not currently make the sort of fine-grained syntactic and semantic judgments needed for automatic evaluation of generation systems. We ran a set of permutations of the words in example (1a) through a number of grammar checkers, including the Microsoft Word, Grammar Expert Plus, Conexor True Styler, Grammar Station, Grammar Slammer, WGrammar and Grammatica grammar checkers. None of the errors in the sentences were identified.

Tree-based evaluation metrics (*e.g.* [14]), while not encoding an explicit model of constituency, can be indirect models of constituency. The task then becomes one of annotating reference and candidate sentences with syntax trees. For evaluation of

general-purpose surface realizers, a treebank can be used; for evaluation of domain-specific surface realizers or selecting a variation from a two-stage surface realizer at run time, this is not currently possible.

The output of a chunker is a shallow model of constituency, is easier to obtain than a full parse tree, and may serve as an approximation to a parse tree for evaluation purposes. To test this, we chunked the sentences in our data using the ILK chunker [18], which chunks noun phrases and prepositional phrases. We used a chunk-based version of simple string accuracy to evaluate the paraphrases. This metric is somewhat correlated with human judgments of adequacy (0.461) and negatively correlated with human judgments of fluency (-0.383). The disagreements are due to two factors. The most frequent is word choice variation; there are also some generated sentences that are much shorter than the original. Performance could perhaps be improved with the inclusion of automatic semantic role labeling.

This method gives the second highest correlation with human judgments of adequacy that we have observed. We therefore recommend that *tree- or chunk-based metrics should be preferred over string-based ones for evaluating adequacy*. However, these metrics do not show promise for evaluating fluency in the absence of a model of word choice variation, or at least the use of multiple reference sentences.

Proposal 3: Models of semantic similarity Existing automatic evaluation methods for generation do not incorporate any measure of semantic similarity other than string equality on words. This affects the evaluation of systems that permit word choice variation, and also those that permit word order variation, since some word order variations (e.g. the use of passive voice) affect word choice.

We have explored two possible solutions to this problem. One can extend existing automatic evaluation metrics like Melamed's F measure using a resource like WordNet, so that the replacement of a word with one of its synonyms is not penalized. This method could not be used for the baseline system data which was created using WordNet. However, we applied this method to the MSA sentences; it performed worst of all the automatic evaluation metrics because it was far too forgiving.

The problem of word choice variation also motivated our decision to include LSA in our experiment. LSA performed well compared to other evaluation metrics. We therefore recommend that *a measure of semantic similarity (e.g. LSA) should be incorporated in automatic evaluation metrics for systems that permit word choice variation*. However, LSA works best when the items being compared are of similar length and are not too short; a single phrase or even a sentence may be too short. We are currently exploring ways to combine semantic similarity and chunk-based metrics.

Word choice variation presents significant difficulty for automatic evaluation of surface realizers, and requires considerable further research.

Proposal 4: Separating different features Recall that our definition of a valid sentence is one that is fluent, adequate and readable. As the experiment in this paper shows, evaluation metrics that are adequate for evaluating adequacy may fail at evaluating fluency and readability.

We propose that the evaluation of surface realization quality should involve more careful analysis than has been previously used, particularly if the surface realizer per-

mits word choice or word order variation. In particular, we recommend that *researchers should evaluate the adequacy, fluency and readability of generator output separately* until there is a metric that can evaluate all three together with high accuracy. Existing string- or tree-based metrics can be used to evaluate adequacy. We recommend the use of multiple reference sentences and tree- or chunk-based metrics where possible. Existing metrics cannot be used to evaluate fluency (at least where there is only one reference sentence), and there is no existing automatic metric that can evaluate readability.

Of course, evaluation of the quality of surface realization output should usually be combined with evaluation of coverage (as in [16]).

7 Conclusions

In this paper, we compared several automatic evaluation metrics, some of which have not previously been used to evaluate the quality of generation system output. We looked at the particular question of whether these automatic evaluation metrics are useful for evaluating the adequacy and fluency of the output of surface realizers that permit variation. We found that these automatic evaluation metrics are not adequate for the task of evaluating fluency, and are only barely adequate for evaluating adequacy, in the context of variation generation. We made several proposals for overcoming this problem, including: use multiple reference sentences, use tree- or chunk-based metrics that give better models of constituent movement, and evaluate adequacy, fluency and readability separately.

This experiment shows that, when selecting an evaluation metric, it is important to consider whether the metric can evaluate the phenomena that the system was designed to handle. There is no single evaluation metric that will work for all surface realizers, across domain, task and discourse type. This makes it harder to compare different surface realizers, but if they perform different tasks it is not clear what use a comparison would be in any case. It is crucial to have a clear understanding of the focus of both surface realizer and evaluation metric before evaluation.

In future work, we plan to explore whether it is possible to use automatic clustering approaches such as those used by [6], together with a Web search engine, to automatically locate multiple reference sentences given a single reference sentence. We also plan to explore other means for automatically evaluating the fluency and readability of generated sentences.

8 Acknowledgments

We thank Dr. Lee and Dr. Barzilay for sharing their data, our judges for their help, and NIST and Dr. Melamed for providing evaluation software. We would like to thank the anonymous reviewers of this paper for their comments. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under Contract No. NBCHD030010. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

References

1. Elhadad, M., Robin, J.: Controlling content realization with functional unification grammar. In: Proceedings of the 6th International Workshop on Natural Language Generation. (1992)
2. Bangalore, S., Rambow, O.: Exploiting a probabilistic hierarchical model for generation. In: Proceedings of COLING 2000. (2000)
3. Langkilde, I.: Forest-based statistical sentence generation. In: Proceedings of ANLP 2000. (2000)
4. McKeown, K.: Paraphrasing using given and new information in a question-answer system. In: Proceedings of ACL 1979. (1979)
5. Murata, M., Isahara, H.: Universal model for paraphrasing – using transformation based on a defined criteria. In: Proccedings of the NLPRS 2001 workshop on Automatic Paraphrasing: Theories and Applications. (2001)
6. Barzilay, R., Lee, L.: Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In: Proceedings of HLT-NAACL 2003. (2003)
7. Barzilay, R., K.McKeown: Extracting paraphrases from a parallel corpus. In: Proceedings of ACL/EACL 2001. (2001)
8. Ibrahim, A., Katz, B., Lin, J.: Extracting structural paraphrases from aligned corpora. In: Proceedings of the 2nd International Workshop on Paraphrasing. (2003)
9. Pang, B., Knight, K., Marcu, D.: Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In: Proceedings of HLT-NAACL 2003. (2003)
10. Shinyama, Y., Sekine, S., Sudo, K., Grishman, R.: Automatic paraphrase acquisition from news articles. In: Proceedings of HLT-NAACL 2002. (2002)
11. NIST: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics (2002)
12. Papenini, K., Roukos, S., Ward, T., Zhu, W.: BLEU: A method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), Thomas J. Watson Research Center, IBM Research Division (2001)
13. Turian, J., Shen, L., Melamed, I.D.: Evalaution of machine translation and its evaluation. In: Proceedings of MT Summit IX. (2003)
14. Bangalore, S., Rambow, O., Whittaker, S.: Evaluation metrics for generation. In: Proceedings of INLG 2000. (2000)
15. Langkilde, I.: An empirical verification of coverage and correctness for a general-purpose sentence generator. In: Proceedings of INLG 2002. (2002)
16. Callaway, C.: Evaluating coverage for large symbolic NLG grammars. In: Proceedings of IJCAI 2003. (2003)
17. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science **41** (1990)
18. Daelemans, W., Buchholz, S., Veenstra, J.: Memory-based shallow parsing. In: Proceedings of CoNLL-99. (1999)