Some Issues in Automatic Evaluation of English-Hindi MT: *More Blues for BLEU*

Ananthakrishnan R[†±], Pushpak Bhattacharyya[†], M Sasikumar[±], Ritesh M Shah[±]

[†]Department of Computer Science and Engineering
Indian Institute of Technology
Powai, Mumbai-400076
India
{anand,pb}@cse.iitb.ac.in

[±] Centre for Development of Advanced Computing (formerly NCST)
Gulmohar Cross Road No. 9
Juhu, Mumbai-400049
India
{sasi,ritesh}@cdacmumbai.in

Abstract

Evaluation of Machine Translation (MT) has historically proven to be a very difficult exercise. In recent times, automatic evaluation methods have become popular. Most prominent among these is BLEU, which is a metric based on *n*-gram co-occurrence. In this paper, we argue that BLEU is not appropriate for the evaluation of systems that produce *indicative* (rough) translations. We use particular divergence phenomena in English-Hindi MT to illustrate various aspects of translation that are not modeled well by BLEU. We show that the simplistic *n*-gram matching technique of BLEU is often incapable of differentiating between acceptable and unacceptable translations.

1 Introduction

Evaluation of Machine Translation (MT) has historically proven to be a very difficult exercise. The difficulty stems primarily from the fact that translation is more an art than a science; most sentences can be translated in many acceptable ways. Consequently, there is no *gold standard* against which a translation can be evaluated.

Traditionally, MT evaluation has been performed by human judges. This process, however, is time-consuming and highly subjective. The investment in MT research and development being what it is, the need for quick, objective, and reusable methods of evaluation can hardly be over-emphasized. To this end, several methods for automatic evaluation have been proposed in recent years, some of which have been accepted readily by the MT community. Especially popular is BLEU, a metric that is now being used in MT evaluation forums to compare various MT systems (e.g., NIST, 2006) and also to demonstrate improvements in translation quality due to specific changes made to systems (e.g., Koehn et al., 2003). BLEU is an *n*-gram co-occurrence based measure – by this we mean that the intrinsic quality of MT output is judged by comparing its *n*-grams with *reference* translations by humans.

Despite its widespread use, there are reservations being expressed in several quarters regarding the simple-mindedness of the measure. Questions have been raised about whether an increase in BLEU score is a necessary or sufficient indicator of improvement in MT quality. It has been argued that while BLEU and other such automatic techniques are useful, they are not a panacea, and that they must be used with greater caution; there is definitely a need to establish which uses of BLEU are appropriate and which are not.

In this paper, we call attention to one specific "inappropriate use" of BLEU for the case of English to Hindi *indicative* translation. Indicative translations – often termed *rough* or *draft-quality* translations – are produced for assimilation rather than dissemination. Given the present state of MT technology, virtually all fully-automatic, general-purpose MT systems can be said to produce indicative translations. Such systems produce understandable output, but compromise on the fluency or naturalness of the translation in the interest of making system development feasible. We use particular divergence phenomena in English-Hindi MT to illustrate various aspects of translation that are not modeled well by BLEU.

Being the most popular of the automatic evaluation techniques, BLEU has served as the means of illustration for most critiques on this topic, and so it is in this paper. However, some of the issues raised are general in nature, and apply to other automatic evaluation methods too.

The paper is organized as follows: We set the background in section 2 by discussing some general issues in MT evaluation. Section 3 contains a brief recap of BLEU. Section 4 reviews and summarizes the existing criticisms of BLEU, and section 5 furthers the argument against BLEU by illustrating how it fails in the evaluation of typical indicative translations. Section 6 concludes the paper and raises questions for further research.

2 Issues in MT Evaluation

For different people concerned with MT, evaluation is an issue in different ways. Potential end-users may wish to know which of two MT systems is better. Developers may wish to know whether the latest changes they have applied to the system have made it better or worse.

At the first level, MT evaluation techniques can be classified as *black-box* or *glass-box*. Black-box techniques consider only the output of the system, whereas glass-box techniques look at the internal components of the system and the intermediate outputs. Glass-box techniques provide information about where the system is going wrong and in what specific way, and are generally part of the developer's internal evaluation of the system.

Evaluation methods (Arnold et al., 1993; White, 2003) can also be (i) *operational* – how much savings in time or cost an MT system brings to a process or application, (ii) *declarative* – how much of the source is conveyed by the translation (fidelity) and how readable it is (intelligibility), or (iii) *typological* – what linguistic phenomena are handled by the system. Operational and declarative methods are by definition of the black-box kind, while typological methods may evaluate both intermediate and final outputs.

BLEU is a declarative evaluation method that provides a score that is said to reflect the quality of the translation. Fidelity and intelligibility are combined in the same score. Declarative methods have been used extensively in MT evaluation, because they are relatively cheap and they measure something that is fundamental to the translation – its quality. This allows a third-party to conduct an evaluation of various systems and publish understandable results.

A) Perfect: no problems in both information and grammar
B) Fair: easy-to-understand with some unimportant information missing or flawed grammar
C) Acceptable: broken but understandable with effort
D) Nonsense: important information has been translated incorrectly

Fig. 1: Example scale for human evaluation of MT

However, declarative evaluation is highly subjective; it is difficult, even amongst translators, to reach a consensus about the best or perfect translation for any but the simplest of sentences. This makes it very difficult to come up with an objective measure of the fidelity and intelligibility of a *candidate* translation. Human ratings (see Fig. 1) have been in use for a long time. Recently, automatic methods have been proposed for this traditionally difficult problem. These techniques compare the candidate translation with one or more *reference* human translations to arrive at a numeric measure of MT quality. The advantages of automatic evaluation are obvious – speed and reusability.

Automatic evaluation techniques have been in use in other areas of natural language processing for some time now. Word Error Rate (Zue et al., 1996) and precision-recall based measures are common in evaluation of speech recognition and spell checking respectively. These measures are also based on comparison with a set of "good" outputs. However, for MT, this kind of evaluation poses some problems: (i) different kinds of quality are appropriate for different MT systems (dissemination vs. assimilation), (ii) different types of systems may produce very different kinds of translation (statistical phrase-based or example-based vs. rule-based), and (iii) the notion of a "good" translation is very different for humans and MT systems.

To see that goodness of translation must be defined differently for humans and MT systems, we note that a human translation, while being faithful to the source, is expected to be clear and unambiguous in the target language. Also, it is expected to convey the same "feel" that the source language text conveys. Consider the following examples of cases where this is especially difficult to achieve: (i) no precise target language equivalent: it is difficult to translate "मेरी दोस्त"

to English without possibly going too far ("my girlfriend") or seeming to over-elaborate the point ("my friend who is a girl" or "my female friend"); (ii) cultural differences: translating "give us this day our daily bread" for a culture where bread is not the staple.

Even the best MT systems of today cannot be expected to handle such phenomena. It is accepted that for unrestricted texts, fully-automatic and human-quality translation is not achievable in the foreseeable future. The compromise is either to produce indicative translations or to use human-assistance for post-editing. Even post-edited output is thought to be inferior to pure human translations, because there is a tendency to post-edit only up to the point where an acceptable translation is realized (Arnold et al., 1993). Thus, a vast majority of MT systems produce translations that are far short of human translations, at least from the viewpoint of stylistic correctness or naturalness.

Such being the situation, the following questions come to mind immediately:

- Can arbitrary systems be pitted against one another on the basis of comparison with human translations? For instance, is it sensible to compare a statistical MT system with a rule-based system, or to compare a system that produces high-quality translations for a limited sub-language with a general-purpose system that produces indicative translations?

- Is it wise to track the progress of a system by comparing its output with human translations when the goal of the system itself cannot be human-quality translation?

In essence, the concern is whether the "failure of MT" (defined using any measure) is simply "failure in relation to inappropriate goals" (translating like a human).

We contend in sections 4 and 5 that the answer to the above questions is "no". But first, a quick recap of BLEU.

3 **BLEU: a recap** (Papineni et al. 2001)

BLEU (BiLingual Evaluation Understudy) evaluates *candidate* translations produced by an MT system by comparing them with human *reference* translations. The central idea is that the more *n*-grams a candidate translation shares with the reference translation, the better it is.

To calculate the BLEU score for a particular MT system, first we need to create a test-suite of sentences in the source language. For each sentence in the suite, we are required to provide one or more high-quality reference translations. Legitimate variation in the translations (word-choice and phrase order) is captured by providing multiple reference translations for each test sentence.

To measure the extent of match between candidate translations produced by the system and reference translations, BLEU uses a modified precision score defined as:

$$p_n = \frac{\sum_{C\in\{Candidates\}} \sum_{n-gram\in C} Count_{clip}(n-\text{gram})}{\sum_{C\in\{Candidates\}} \sum_{n-gram\in C} Count(n-\text{gram})},$$

where C runs over the entire set of candidate translations, and $Count_{clip}$ returns the number of n-grams that match in the reference translations.

Having no notion of recall, BLEU needs to compensate for the possibility of proposing high-precision translations that are too short. To this end, a brevity penalty is introduced:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-\frac{r}{c}} & \text{if c} \le r \end{cases},$$

where c is the cumulative length of the set of candidate translations and r, that of the set of reference translations.

Finally, the BLEU score is calculated as:

$$\log BLEU = \min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n \log p_n ,$$

where $N = 4$ (unigrams, bigrams, trigrams, and 4-grams are matched) and $w_n = N^{-1}$ (n-grams of all sizes have the same weight).

(Papineni et al. 2001) and (Doddington, 2002) report experiments where BLEU correlates well with human judgments.

4 Criticisms of BLEU

Notwithstanding its widespread use, there have been criticisms of BLEU, most significant among these being that it may not correlate well with human judgments in all scenarios. We review these criticisms in this section.

1. **Intrinsically meaningless score**: The first criticism of BLEU is that the score that it provides is not meaningful in itself, unlike, say, a human judgment or a precision-recall score. It is useful only when we wish to compare two sets of translations (by two different MT systems or by the same system at different points in time). Newer evaluation measures have attempted to address this problem (Akiba et al., 2001; Akiba et al., 2003; Melamed et al., 2003).

2. **Admits too much variation**: Another criticism is that the *n*-gram matching technique is naïve, allowing just too much variation. There are typically thousands of variations on a hypothesis translation – a vast majority of them both semantically and syntactically incorrect – that receive the same BLEU score. Callison-Burch et al. (2006) note that "phrases that are bracketed by bigram mismatch sites can be freely permuted, because reordering a hypothesis translation at these points will not reduce the number of matching *n*-grams and thus will not reduce the overall BLEU score."

3. **Admits too little variation:** Languages allow a great deal of variety in choice of vocabulary. BLEU, on the other hand, treats synonyms as different words. Word choice is captured only to a limited extent even if multiple references are used. Uchimoto et al. (2005) propose a measure which matches word classes rather than words.

4. **An anomaly – more references do not help:** It was claimed originally that the more reference translations per test sentence, the better. However, the NIST evaluation (Doddington, 2002) and Turian et al. (2003) report that the best correlation with human judgments was found with just a single reference translation per test sentence. This goes entirely against the rationale behind having multiple references – capturing natural variation in word choice and phrase construction. No convincing explanation has been found for this yet.

5. **Poor correlation with human judgments**: The final, and the most damning, criticism is that BLEU scores do not correlate with human judgments generally. Turian et al. (2003) report experiments showing that the correlation estimates on shorter documents are inflated – with larger corpora the correlation between BLEU and human judgments is poor. (Carlson-Burch et al., 2006) compares BLEU's correlation with various SMT systems and a rule-based system (Systran), again with discouraging results.

The main point that comes out of these criticisms is that BLEU needs to be used with caution; there is a need for greater understanding of which uses of BLEU are appropriate, and which are not. (Calrson-Burch et al., 2006) suggests that it is not advisable to use BLEU for comparing systems that employ different strategies (comparing phrase-based statistical MT systems with rule-based systems, for example). It is also suggested that while tracking broad changes within a single system is appropriate, the changes should be those aspects of translation that are modeled well by BLEU. However, the question as to what aspects of translation are not modeled well by BLEU has not been addressed so far. We believe that this question needs to be looked at in more detail, and we make a beginning in this paper. Previous criticisms have argued against BLEU based either on hypothetical considerations (phrase permutations that BLEU allows) or on its performance on large test-sets; we supplement these criticisms by characterizing BLEU's failings in terms of actual issues in translation.

5 Evaluating Indicative Translations: where BLEU fails

We now proceed to look at specific phenomena that occur in English-to-Hindi indicative translation, which cause BLEU to fail. The results suggest that automatic evaluation techniques like BLEU are not appropriate in cases where the MT system's output is meant just for assimilation, and is often, by intention, not as natural as human translations.

5.1 Indicative translation: a representative characterization

As mentioned in section 2, MT is a difficult problem, more so for widely divergent language pairs such as English-Hindi. To achieve fully-automatic MT for unrestricted texts, developers have to compromise on the quality of the translation – the goal in such scenarios is indicative rather than perfect translation. Indicative translations are understandable but often not very fluent in the target language.

In this context, we look at a one possible characterization of indicative translation: Consider a system that performs the following basic steps in English to Hindi transfer (Rao et al., 1998):

- Structural transfer: this involves (i) changing the Subject-Verb-Object (SVO) order to Subject-Object-Verb (SOV), and (ii) converting post-modifiers to pre-modifiers
- Lexical transfer: this involves (i) looking up the appropriate equivalent for the source language word in a transfer lexicon (may require WSD), (ii) inflecting the words according to gender, number, person, tense, aspect, modality, and voice, and (iii) adding appropriate case-markers.

We think of this as a system that produces indicative translations. Now, we look at certain divergence phenomena between English and Hindi (Dave et al., 2002) that are not dealt with adequately by such a system. We do not claim that all these phenomena are impossible to handle, only that the processing involved is beyond the basic steps listed above and represents progress from indicative to human-quality translation. For a system aiming for indicative translation, there are certain divergence phenomena that have to be handled to keep translations from dropping below the acceptable level, and certain others that may be ignored while still keeping the translations understandable. We would expect any evaluation mechanism for such an MT system to make this difference. Below, we illustrate divergence phenomena between indicative and human translations where BLEU's judgment is contrary to what is expected – in some cases, acceptable translations are penalized heavily, and in others, intolerable translations escape with very mild punishment indeed.

5.2 Categorial divergence

Indicative translation is often unnatural when the lexical category of a word has to be changed during translation. In the following example, the verb-adjective combination *feeling hungry* in the source language (**E**) is expressed in the human reference translation (**H**) as a noun-verb combination ("भूख लगना"), whereas this change does not occur in the indicative translation.

Though **I** (the candidate indicative translation) is easily understandable, the BLEU score is 0, because there are no matching *n*-grams in **H.**

E: I am feeling hungry
H: मुझे भूख लग रही है
to-me hunger feeling is
I: मैं भूखा महसूस कर रहा हूँ
I hungry feel doing am

***n*-gram matches**: unigrams: 0/6; bigrams: 0/5; trigrams: 0/4; 4-grams: 0/3

We have quoted the precision of *n*-gram matching for all examples, because, as mentioned earlier, the BLEU score by itself does not reveal much and is useful only in comparison. In the above example, unigram precision is 0 out of 6, bigram precision is 0 out of 5, and so on.

5.3 Relation between words in noun-noun compounds

The relation between words in a noun-noun compound often has to be made explicit in Hindi. For example, *cancer treatment* becomes "कैनसर का इलाज" (*treatment of cancer*) whereas *herb treatment* is "जड़ी-बूटियों द्वारा/ से इलाज" (*treatment using herbs* and not *treatment of herbs*). In the following example, we have a five word noun chunk (*ten best Aamir Khan performances*). The indicative translation follows the English order, again leading to an understandable translation, but a low BLEU score, with none of the higher-order *n*-grams matching.

E: The ten best Aamir Khan performances
H: आमिर ख़ान की दस सर्वोत्तम पर्फ़ार्मन्सस
Aamir Khan of ten best performances
I: दस सर्वोत्तम आमिर ख़ान पर्फ़ार्मन्सस
Ten best Aamir Khan performances

***n*-gram matches**: unigrams: 5/5; bigrams: 2/4; trigrams: 0/3; 4-grams: 0/2

5.4 Lexical divergence: beyond lexicon lookup

In the translation of expressions that are idiomatic to a language, target language words are not literal translations of the source language words. Such translation is beyond the purview of MT systems. The following is such an example where the drop in BLEU score is unwarranted.

E: Food, clothing and shelter are a man's basic needs

H: रोटी, कपड़ा और मकान एक मनुष्य की

bread clothing and house a man of

बुनियादी ज़रूरतें हैं

basic needs are

I: खाना, कपड़ा, और आश्रय एक मनुष्य की

food clothing and shelter a man of

बुनियादी ज़रूरतें हैं

basic needs are

***n*-gram matches**: unigrams: 8/10; bigrams: 6/9; trigrams: 4/8; 4-grams: 3/7

Non-literal translation also happens due to cultural differences, such as when translating the expression *bread and butter*, which could be translated as “रोज़ी-रोटी” (*livelihood-bread*), “दाल-रोटी” (*dal-bread*), or “रोटी और मक्खन” (*bread and butter*) in different contexts.

5.5 Pleonastic divergence

In the following sentence, the word *it* has no semantic content (such a constituent is called a pleonastic). The indicative translation is objectionable, but the number of *n*-gram matches is high, including several higher order matches.

E: It is raining

H: बारिश हो रही है

rain happening is

I: यह बारिश हो रही है

it rain happening is

***n*-gram matches**: unigrams: 4/5; bigrams: 3/4; trigrams:2/3; 4-grams: 1/2

5.6 Other stylistic differences

There are also other stylistic differences between English and Hindi. In the following example, the transitive verb in English maps to an intransitive verb in Hindi. The sentence should be translated as “*In the Lok Sabha, there are 545 members.*” The indicative translation clearly conveys an incorrect meaning, but the number of *n*-gram matches is still quite high.

E: The Lok Sabha has 545 members

H: लोक सभा में ५४५ सदस्य हैं

Lok Sabha in 545 members are

I: लोक सभा के पास ५४५ सदस्य हैं

Lok Sabha has/near 545 members are

***n*-gram matches**: unigrams: 5/7; bigrams:3/6; trigrams: 1/5; 4-grams: 0/4

5.7 WSD errors and transliteration

As mentioned in section 4, words in the candidate translation that do not occur in any reference translation can be replaced by any arbitrary word. Consider the following example:

E: I purchased a bat

H: मैने एक बल्ला खरीदा (reference)

I a cricket-bat bought

I: मैने एक चमगादड़ खरीदा

I a bat (mammal) bought

***n*-gram matches**: unigrams: 3/4; bigrams: 1/3; trigrams:0/2; 4-grams: 0/1

Now, in cases where the lexicon does not contain a particular word, most MT systems would use transliteration as in the following:

I: मैने एक बैट खरीदा

I a bat (transliteration) bought

This translation would receive the same BLEU score as the translation with the WSD error, which is clearly ridiculous.

5.8 Discussion

Table 1 puts together the average precision figures (P) for the examples cited in this section. P is the mean of the modified precision (p_n) of unigrams, bigrams, trigrams and 4-grams:

$$P = \frac{\sum_{n=1}^{4} p_n}{4}$$

Though the exact precision figures are not very significant, as we are dealing with particular examples, what is important to note is that in each case BLEU’s model of the variation allowed by the target language (indicative Hindi) is flawed. The acceptable translations demonstrate variations that are allowed by the target language, but not allowed by BLEU – these variations cannot be captured simply by increasing the number of reference translations, because native speakers of Hindi can never be expected to produce such constructs. On the other hand, the unacceptable translations demonstrate variations not allowed by the target language that, however, are allowed by BLEU.

Divergence or problem example	**Average BLEU precision**	**Translation acceptable?**
Categorial (5.2)	0	Yes
Noun-noun compounds (5.3)	0.38	Yes
Lexical (5.4)	0.6	Yes
Pleonastic (5.5)	0.68	No
Stylistic (5.6)	0.35	No
WSD error (5.7)	0.27	No
Transliteration (5.7)	0.27	Yes

Table 1: Summary of examples

As mentioned earlier, the problems and divergence phenomena that we have discussed in this section are representative of what a typical English-Hindi MT system would need to address to move towards human-quality translation. However, some of these phenomena may be ignored when the objective is simply indicative translation – this can lead to substantial savings in the time and cost required for system development. Indeed, at the present stage of research in English-Hindi MT, it may even be necessary to ignore some of these phenomena to make MT feasible.

In this situation, it is imperative that the strategy used for evaluation models the indicative MT task. The gradation in evaluation should be in sync with the standards that are set forth as the objective of the system. The issues raised in this section suggest that BLEU fails on this count – using BLEU for comparing or tracking the progress of such a system is likely to be misleading.

6 Conclusion

In this paper, we have reviewed existing criticisms of BLEU and examined how BLEU fares in the judgment of various divergence phenomena between indicative and human translations for English-Hindi MT. What we have shown is that evaluation using BLEU is often misleading – BLEU overestimates the importance of certain phenomena and grossly underestimates the importance of others. The broader concern, which has significance even beyond indicative MT, is that BLEU is unable to weed out structures and word choices that make the translation absolutely unacceptable.

From the point of view of indicative translation, MT researchers and developers would be expected to tackle those problems first that would affect the understandability and acceptability of the translation. Fluency of translation is often intentionally sacrificed to make system development feasible. Engineering such a system requires a developer to make many choices regarding the importance of handling various phenomena based on a deep knowledge of the idiosyncrasies of the languages involved. Ideally, the evaluation method used for such a system also should factor in these choices. At any rate, the method must be able to grade translations according to the standards set forth for the system. The issues raised in this paper suggest that BLEU, in its current simplistic form, is not capable of this. Our contention, based on this initial study, is that BLEU is not an appropriate evaluation method for MT systems that produce indicative translations. To further substantiate this claim, we are working on creating larger test-sets of sentences exhibiting each of the divergence phenomena discussed in section 5.

Can BLEU be adapted for evaluation of such systems, possibly, by modifying the matching strategy or the reference sets to allow specific features that occur in indicative translations? The difficulty with this is that the nature of indicative translations would vary across systems and over time. Thus, we are faced with the problem of measuring against a benchmark that is itself unstable. Moreover, any such changes to BLEU are likely to compromise on its simplicity and reusability – characteristics that have made it the evaluation method of choice in the MT community.

Another important question is whether BLEU is a suitable evaluation method for “into-Hindi” MT systems? How do the free word-order, case-markers, and morphological richness of Hindi affect the *n*-gram matching strategy of BLEU?

Finally, how far do the concerns raised in this paper regarding BLEU apply to other automatic measures, such as word error rate and edit distance-based measures?

Further theoretical and empirical work is required to answer these questions fully. Meanwhile, it might be advisable not to be overly reliant on BLEU, and allow it to be what its name suggests: an "evaluation understudy" to human judges.

Acknowledgements

Our thanks to Jayprasad Hegde for his feedback.

References

Y. Akiba, K. Imamura, and E. Sumita. 2001. Using multiple edit distances to automatically rank machine translation output. In *Proceedings of MT Summit VIII*, pages 15–20.

Y. Akiba, E. Sumita, H. Nakaiwa, S. Yamamoto, and H.G. Okuno. 2003. Experimental comparison of MT evaluation methods: RED vs. BLEU. In *Proceedings of MT Summit IX*.

Doug Arnold, Louisa Sadler, and R. Lee Humphreys. 1993. Evaluation: an assessment. *Machine Translation*, Volume 8, pages 1–27.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research, In *Proceedings of the EACL*.

Shachi Dave, Jignashu Parikh, and Pushpak Bhattacharyya. 2002. Interlingua based English Hindi machine translation and language divergence, *Journal of Machine Translation (JMT)*, Volume 17.

Doddington, G. 2002. Automatic evaluation of machine translation quality using *n*-gram co-occurrence statistics, In *Proceedings of the Second International Conference on Human Language Technology*.

Philipp Koehn and Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *Proceedings of the Human Language Technology Conference (HLT)*, pages 61–63.

NIST. 2006. The 2006 NIST machine translation evaluation plan (MT06). http://www.nist.gov/speech/tests/mt/doc/mt06_evalplan.v4.pdf

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation, IBM research report rc22176 (w0109-022). Technical report, IBM Research Division, Thomas, J. Watson Research Center.

Rao D., Bhattacharya P., and Mamidi R. 1998. Natural language generation for English to Hindi human-aided machine translation. In *Proceedings of the International Conference on Knowledge Based Computer Systems (KBCS).*

E. Sumita, S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishikawa, and S. Shirai. 1999. Solutions to problems inherent in spoken-language translation: the atr-matrix approach. In *Proceedings of MT Summit VII*, pages 229–235.

J. Turian, L. Shen, and I.Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. In *Proceedings of MT Summit IX*.

Kiyotaka Uchimoto, Naoko Hayashida, Toru Ishida, and Hitoshi Isahara. 2005. Automatic Rating of Machine Translatability. In *Proceedings of MT Summit X*.

John S. White. 2003. How to evaluate machine translation. *Computers and Translation*, Harold Somers (Ed.). John Benjamins Publishing Company.

Victor Zue, Ron Cole, and Wayne Ward. 1996. *Survey of the State of the Art in Human Language Technology*, chapter 1, section 2, Speech Recognition.