RC22796 (W0305-050) May 9, 2003
Computer Science

IBM Research Report

A Fine-grained Evaluation Framework for Machine Translation System Development

Nelson Correa
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

A Fine-grained Evaluation Framework for Machine Translation System Development

Nelson Correa
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
necorrea@us.ibm.com

Abstract

Intelligibility and fidelity are the two key notions in machine translation system evaluation, but do not always provide enough information for system development. Detailed information about the type and number of errors of each type that a translation system makes is important for diagnosing the system, evaluating the translation approach, and allocating development resources. In this paper, we present a fine-grained machine translation evaluation framework that, in addition to the notions of *intelligibility* and *fidelity*, includes a typology of errors common in automatic translation, as well as several other properties of source and translated texts. The proposed framework is informative, sensitive, and relatively inexpensive to apply, to diagnose and quantify the types and likely sources of translation error. The proposed fine-grained framework has been used in two evaluation experiments on the LMT English-Spanish machine translation system, and has already suggested one important architectural improvement of the system.

1 Introduction

Most current approaches to machine translation (MT) system evaluation focus exclusively on the translation quality notions of *intelligibility* of the translated text and *fidelity* of the translated text to the source text,[1] or use extensive taxonomies that seek to accommodate the evaluation needs of multiple kinds of users in a variety of translation task contexts (EAGLES, 1996; Hovy *et al.*, 2002).

The existing approaches may thus be viewed as either general and not particularly intended for routine use in MT system development, or not very informative regarding the strengths and weaknesses in a given system's performance. As Reeder (2001) puts it, "*. . . the measurements have failed to meet the desired properties of replicability, scalability and informativeness for users and developers.*"

In this paper we present a fine-grained evaluation framework to improve that situation, particularly for MT system development. The new framework assigns a multi-featured *evaluation vector* $\mathcal{V}$ to translation segments, documents, and ultimately the MT system under evaluation. In addition to the quality-related features of fidelity and intelligibility, $\mathcal{V}$ also includes features for the types of errors most common in the translation task of interest, as well as several properties of the source and translated texts.

The motivation for this work arose out of the need to evaluate and document the performance of the LMT English-Spanish (LMTES) machine translation system (McCord, 1989; McCord and Bernth, 1998), a transfer-based system whose main application is Web page translation.

While there is a wide range of evaluation features that could be selected (cf. the feature set of the ISLE framework (ISLE, 2000)), we take the opportunistic approach of Gdaniec (1999), and select features that are informative and relevant for system development, and a balance between the number of features selected, their ease of measurement, and the overall evaluation effort required.

Two particular evaluation features of the framework, selected given our experience with the LMTES system, are the number of translation errors attributable to incorrect target lexical selections (*TL*) and incorrect named entity translations (*NE*).

[1] The related notions of *fluency* and *adequacy* are also commonly used; they are particular measures of *intelligibility* and *fidelity*, respectively; see for example (White, 1995; LDC, 2002; Doddington, 2001).

The framework currently assumes manual feature evaluation using a prototype Web application, which provides access to the source and target segments, and optionally to the source syntactic analysis trees. Despite the larger number of features involved, our experience shows that the effort required for fine-grained evaluations is only marginally higher than the effort required to produce the translation quality score alone (fidelity plus intelligibility). Furthermore, each of the evaluation components is susceptible to automation; cf. Papineni (2002) for automation of the quality score.

We have conducted two evaluation experiments using the framework, involving approximately 1,500 segments and 21,000 words. The results have provided concrete development information, guiding the areas where development work should be concentrated, and has already suggested one important architectural improvement of the system.

The paper is organized as follows: Section 2 presents relevant previous work on MT evaluation; Section 3 presents the fine-grained evaluation framework; in Section 4 we present the results of applying the new framework to the evaluation of the LMTES translation system; lastly, in Section 5 we discuss some extensions of this work, including options for evaluation automation.

2 Previous Work on MT Evaluation

MT evaluation (MTE) has been an active field, prompting some to suggest that MTE is better understood than MT itself (Wilks, 1994). Most pragmatic approaches use primarily the translation quality notions of *fidelity* and *intelligibility* of the translated text. That is especially true of the approaches used in regularly scheduled MT evaluations, such as (LDC, 2002) recently, and the DARPA MT evaluations, from the mid-1990s (White, 1995).

European and U.S. agencies have funded development of general MTE frameworks, as part of the EAGLES (1996) and ISLE (2000) projects. Those approaches build on previous work on software evaluation standards and take a comprehensive view of MTE: They define two feature taxonomies, one relating context of use to quality characteristics, and the second quality characteristics to evaluation metrics (Hovy, 1999; Hovy *et al.*, 2002).

The general frameworks need to be customized and adapted for each particular use situation, and appear more directed towards one-time system evaluations, than routine evaluation of one system in the course of development work.

The component-based evaluation methodology of Nyberg *et al* (1994) refers to MT system structure and uses fine-grained evaluation features. Their method is "glass-box," assuming access to system-internal representations, and does not address translation errors, such as named entity errors, that have now become prominent and we consider below.

Other recent work has explored the use of simpler, easier-to-measure features and their correlation with quality features like intelligibility and fidelity. For some multi-feature approaches to evaluation, see (Reeder, 2001; Reeder *et al.*, 2001; Vanni and Miller, 2001).

Finally, other recent work has proposed methods and tools for automatic or automated evaluation (Nießen *et al.*, 2000; Papineni *et al.*, 2002).

Our fine-grained evaluation framework detailed in the following section can be seen as one particular adaptation of a general MTE framework such as ISLE for our purpose: MT evaluation that is sensitive and informative for MT system development.

3 Fine-grained Evaluation Framework

Our fine-grained evaluation framework provides an overall translation quality score at the segment, document, and system levels, and concrete information about the type and possible source of translation errors on the system output. The particular set $\mathcal{F}$ of evaluation features adopted reflects our interest in evaluation of LMTES, a transfer-based system whose main application is Web page translation.

3.1 Segment level evaluation components

At the *segment level* we include (i) a translation quality score, as well as features to indicate possible errors in (ii) source tokenization and segmentation, (iii) handing of HTML tag markup, (iv) unknown words, (v) source analysis, (vi) named entity translation, and (vii) lexical selection in the target language, among others. We also found it important to include a feature to indicate (viii) defects in the input

source segments (e.g., misspellings).[2] On the output side, we have features to indicate (ix) the grammaticality and (x) the style of the translated text.

Manual MT quality evaluations are expensive and our framework is not intended to address that.[3] One goal of our framework, however, is that manual evaluation take not much longer than arriving at a translation quality score alone. Thus, to keep the evaluation simple and fast, the type of the other features is either *boolean* or an *integer* count of possible errors.

Table 1 shows the ten features and types selected for evaluation vectors at the segment level.

Feature	Description	Type
Q	***Q***uality score	[0–5]
IN	***IN***put segment error	int
SEG	***SEG***mentation error	bool
TAG	***TAG*** (markup) error	bool
UW	***U***nknown ***W***ord	int
NE	***N***amed ***E***ntity error	int
AN	Source ***AN***alysis error	bool
TL	***T***arget ***L***exical error	int
TG	***T***arget ***G***rammar error	bool
TS	***T***arget ***S***tyle error	bool

Table 1: Segment Evaluation Features

Q: A 0–5 numeric score of segment translation quality, with "0" lowest and "5" highest. This is a combined measure of *Intelligibility* and *Fidelity*, which we adopted from previous LMT evaluations. See the Appendix for the detailed meaning.

IN: The number of defects in the current input segment. This feature is used to track what target errors may be attributed to input defects. Typical input defects are misspellings, bad grammar, etc.

SEG: "1" if there is a tokenization or segmentation problem for this segment; "0" otherwise. The typical problem is the break up of a single sentence into multiple segments.

TAG: "1" if there are (HTML) markup tag errors; "0" otherwise. A typical error is wrong handling of font change of hyperlink tags in the source.

[2] This feature is required to trace what fraction of output errors is caused by input, rather than system error.

[3] Most manual evaluation effort is in arriving at the translation quality score, which alone requires reading the target, and source and/or reference translations.

UW: The number of unknown words in the current segment.

NE: The number of wrong analyses or translations of named entities (people, organizations, places, etc.) in the input segment. A typical error is literal, non-idiomatic translation of a named entity.

AN: "1" if the *source* segment fails to parse (incomplete) or gets an incorrect parse; "0" otherwise. (Currently we focus on incomplete parses; parse correctness is evaluated in few cases only.)

TL: The number of incorrect lexical choices in the translation of the current segment. Lexical transfer problems are typically caused by defects in the bilingual transfer lexicon, but may also be due to other reasons, such as incorrect source analysis.

TG: "1" if the translated segment is ungrammatical in the target language; "0" otherwise. Typical examples include lack of grammatical agreement where required (e.g., Subject-Predicate), repeated words (e.g., double articles), etc. We also include incorrect, literal translation of idiomatic expressions in this category (e.g., "he went broke").

TS: "1" if the translated segment includes grammatical but non-idiomatic material in the target language; "0" otherwise. This includes lexical choices that do not collocate well in the target language, wrong punctuation or capitalization, etc.

From the above error features, we define a boolean *error* feature *E* and an *error count* feature *SE*, useful for document and system level evaluation.

E: "1", if any feature $SEG, \ldots, TS$ is non-zero; "0" otherwise. This field is the logical *or* of the last eight features. Note that *IN* is not included.

SE: The sum of the features $SEG, \ldots TS$.

The features above were pragmatically selected and the feature types have been kept simple to allow for fast and consistent manual evaluation.

Most features assume a "black-box" view of the system and are architecture-independent. The main exception to this is the source analysis feature (*AN*), which requires access to source analysis representations, and furthermore is geared towards a transfer-based system such as LMT. Features *SEG* and *UW* may also require, in a few cases, access to system-internal representations. The other seven features, however, can be evaluated on the basis of the input and output segments alone, and so are "glass-box" and generically applicable to any MT system.

According to Table 1, a translated segment is always assigned a translation quality score and may be marked as having one or more translation errors.

The evaluation feature set is

$$\mathcal{F} = \{\text{Q, IN, SEG, TAG, UW, NE, AN, TL, TG, TS}\} \quad (1)$$

and for each $i \in \mathcal{F}$, we use f_i to denote the feature's value. More generally, we write $f_{i,\,s,\,d,\,sys}$ to denote the value of feature i for segment s in document d, as measured on system sys.

The evaluation vector assigned to each segment s is then

$$\mathcal{V}_s =< f_{i,\,s} : \text{ for } i \in \mathcal{F} > \quad (2)$$

Below we also write $f_{i,\,d,\,sys}$ and $f_{i,\,sys}$ to denote the corresponding feature values at the document and system levels.

3.2 Measurement criteria

The Appendix provides the criteria for assignment of the LMT segment quality score Q. More detailed presentation of the criteria for the other features is beyond the scope of this paper.

3.3 Document level evaluation

At the *document level* the evaluation vector includes the same set of features as the segment level evaluation, plus the following two:

S: The number of segments in the document.

DE: The total number of errors in the document.

The document-level feature values are defined as the mean of the corresponding segment-level features. The total error number *DE* is the sum of the error counts *SE* for the segments in the document:

$$f_{DE,\,d} = \sum_{s=1}^{S} f_{SE,\,s,\,d} \quad (3)$$

The document quality score is taken as the feature Q, which is the arithmetic mean of the segment scores, regardless of segment length.

3.4 System level evaluation

At the *system level* the evaluation vector consists of the features from the document level evaluation, plus the following one:

D: the number of documents used in the system evaluation.

The system-level feature values are defined as the weighted average of the corresponding document-level features, where the weight w_d for document d is given as the number of segments in the document, divided by the total number of segments:

$$f_{i,\,sys} = \sum_{d=1}^{D} w_d \cdot f_{i,\,d,\,sys} \quad (4)$$

where

$$w_d = \frac{f_{S,\,d}}{\sum_{j=1}^{D} f_{S,\,j}} \quad (5)$$

The final system quality score is taken as the feature *Q*, which according to our definition is the weighted average of the document scores.

4 Application of the Framework

We conducted two evaluation experiments on the LMT English-Spanish translation system using the above framework, with a corpus of approximately 1,500 segments and 21,000 words.

The experiments were designed to make an LMTES diagnostic that would be useful for further system development. Indirectly, we also were interested in testing the sensitivity and informativeness of the evaluation framework.

In the first experiment, the selected corpus was translated into Spanish with LMTES, taking as baseline the version of November, 2002. In the second experiment the system was further developed using a small development subset of the first corpus; the performance of the improved system was then measured on test and development corpora.

4.1 Evaluation corpus and system

We selected seven HTML documents from a variety of domains and document styles, with a total of 1,516 segments and 20,946 words. The average document length was 228 segments and the average segment length was 13.8 words. The document sources included two U.S. government agencies, one city agency, one university, one computer company, one news source, and one Web technology reference.

The evaluations were conducted by two Spanish-English bilinguals, with the source, source analysis and translated segments, and with a field for each segment evaluation feature.

To test actual system quality in the translation task selected (Web page translation), we worked with the source HTML documents, rather than pre-segmented text files alone. This allowed us to evaluate errors like segmentation and handling of HTML markup, that wouldn't be evaluated on pre-segmented text files without markup.

4.2 Baseline evaluation experiment

In the first experiment the evaluation was done by one evaluator according to the criteria defined.

The baseline system score was 3.41 on our 0 to 5 quality scale, with a standard deviation of 0.34. The lowest document score was 2.9, and the highest 3.74. The total number of errors identified in the evaluation was 2,278, or an average of 1.5 errors per segment. Of the 1,516 segments evaluated 79%, or 1,192, had some kind of error.

Table 2 shows the percentage of segments per quality score level, 0 to 5, in the baseline system. 22% of the segments had the top score of 5, while the bulk (58%) had a score of 3 or lower, which renders them "informative" but with major translation errors, in what is called a *gisting* application.

Score	Count	%
5	340	22%
4	298	20%
3	489	32%
2	219	14%
1	69	5%
0	101	7%

Table 2: Baseline Translation Quality

Table 3 shows the error type distribution, as the percentage of error segments for each error type.

The error type distribution shows that deficiency in the bilingual transfer lexicon is the largest source of errors. Named entity translation is another significant source.[4] Note that the error sources are interrelated, since a failure, for example, in named entity parsing can lead to other analysis errors, and to target lexical and grammar errors.

[4] LMTES uses multi-word expressions, and subgrammars for numbers and time expressions, for the translation of named entities. We included the *NE* feature in the evaluation framework since named entity errors have a major effect on the intelligibility and fidelity of the translation.

Error Type	Count	%
IN	11	1%
SEG	210	14%
TAG	141	9%
UW	85	6%
NE	208	14%
AN	149	10%
TL	678	45%
TG	611	40%
TS	223	15%

Table 3: Baseline Error Type Distribution

4.3 System development experiment

In the second experiment, we selected 120 segments from the original corpus; in general, we selected the top 20 to 30 segments from five of the source documents. The selected corpus had 1,752 words and an average sentence length of 14.6. The shortest segment was 1 word, and the longest 39.

The goal of this experiment was to seek, in a principled way and with a limited amount of development effort, improvement on the translation of the development corpus, by working separately on the three major stages of the LMT translation process: (1) Source tokenization, segmentation, and analysis, (2) lexical transfer, and (3) syntactic transformation and target output generation.

The improved system was evaluated again on the development corpus and on an independent regression test corpus of 4,239 segments. For this experiment, the evaluations were conducted by three evaluators, fluent English-bilinguals and native Spanish speakers. The evaluator responses were averaged for the integer-type features, and taken as majority vote for the boolean-type features.

The experiment was designed to test the translation quality improvement that could be obtained by working with the development corpus, and to measure the contribution that each stage can make towards quality improvement.

The baseline system score was 3.37 on the 120 segment corpus subset. The score improvement on the development corpus was $+0.83$ (score 4.2), while on the test corpus it was $+0.01$.

Table 4 shows the distribution of segments per

each quality score level, 0 to 5, and the change in the score distribution for the baseline system, the improved system on the development corpus, and the improved system on the test corpus.

Score	Baseline	Development	Test
5	14%	46%	14%
4	14%	29%	15%
3	42%	21%	42%
2	24%	2%	23%
1	6%	1%	6%
0	0%	0%	0%

Table 4: Translation Quality Improvement

There is a significant improvement in translation quality, as measured on the development corpus. The percent of segments with the top score of 5 (high quality) increased from 14% to 46%, while the percent of segments with a score of 3 or lower (*gist* quality) was reduced from 58% to 25%.

The *relevant* system quality improvement (that is, testing on unseen data) obtained in this experiment, however, was positive but not significant. On the regression test corpus only 136 of 4,239 segments changed; 74% of the segment score changes were positive.[5]

Table 5 shows the distribution of segment translation errors, per error type, using the percentage of the total number of segments for each error type.

4.4 Results analysis

Certain types of errors, such as unknown words (*UW*) and named entities (*NE*), at least in a closed development corpus, can easily be eliminated completely or nearly so.

Errors in target lexical selection (*TL*) can also be reduced substantially, e.g., from 60% to 10%, but are much more difficult to eliminate completely. LMT, for example, works on individual sentences and there is no general inter-sentential context for word sense disambiguation, [6] so some words cannot be reliably translated. [7]

Error Type	Baseline	Development
IN	2%	2%
SEG	11%	1%
TAG	6%	3%
UW	11%	0%
NE	19%	1%
AN	23%	3%
TL	60%	10%
TG	59%	29%
TS	37%	42%

Table 5: Error Type Distribution and Improvement

In the development experiment the effort on fully improving the grammatically of the translations (*TG*) was limited, so in that category the reduction in error rate is not as dramatic as with the previous error categories.

Finally, it may seem paradoxical that the number of style errors in the target text (*TS*) was not reduced in the development experiment, but instead increased. The percentage of segments with style error went up from 37% to 42%. That is an artifact of our evaluation criteria, where style errors are not (necessarily) marked on a translated segment if there are much more serious defects in the segment, such as a failed source analysis or altogether a lack of translation. In those cases, no style error is marked.

The larger number of style errors in Table 5 is a reflection of the overall improvement in translation quality, so that it becomes meaningful to evaluate segment translations according to style.

4.5 Web-based MT Evaluation Application

Our framework has been implemented as a Web-based XML application, which can dynamically generate Web forms with the required input boxes from MT system outputs. One major advantage of a

[5] The high score improvement testing on development data and the small improvement on the test data are not surprising, given the very small size of the development corpus. The result shows the difficulty of achieving general MT quality improvements, and the importance of working with large development corpora, focusing on error types with the most impact.

Our interest in this experiment was to test the sensitivity and informativeness of the fine-grained evaluation framework.

[6] There is a mechanism to activate one or more given "subject areas" for an entire text, which can be used to control lexical selection and translation. However, that is not sufficient for full word sense disambiguation.

[7] An example is the sentence "Runners in dark shades and earth tones, that's all anyone wants, ...", where the sense of *"runners"* refers to *rug runners*, but that is clear only from the larger context of the news article.

Web application is that it can be widely accessible, with minimal client software and data requirements; only a recent browser (IE-6) is required.

5 Future work

We plan to explore the correlation between the different subsets of features in our evaluation vectors. This is similar to other work, for example, studying the correlation between translation quality and named entity translation (Reeder *et al.*, 2001).

The detailed LMTES evaluation results of the new framework provide an indication of the areas of development most needed on the translation system, and their relative importance. One such area is improved translation of named entities; in the development experiment the incidence is in 19% of segments, and the score impact is between -1 and -2 points each time. We are addressing that problem in separate work, using a pipelined NLP processing model where named entities are identified and translated in a separate, prior stage to general translation.

Our evaluation also shows that error correction of the input (e.g., spelling and grammar correction) is not an important issue for the present (Web page) translation task; that would be more relevant in other tasks, such as chat or email translation.

An important area of work is automation of the evaluation framework. We are currently incorporating the BLEU measure (Papineni *et al.*, 2002) for estimation of segment and document translation quality. Furthermore, we expect unigram precision to show high correlation with target lexical error (*TL*), and statistical target language models to be useful in assessing target grammar and style (*TG, TS*).

Similarly, the unknown word (*UW*) and source analysis (*AN*) features can be automated using the trace options of LMT. Evaluation of other features could be sped up by using Bernth's (1999) Translation Confidence Index and selecting for human evaluation only segments with low confidence.

Finally, we expect that the set $\mathcal{F}$ of evaluation features in our framework will be modified as needed to fit the characteristics of the system under evaluation and the translation task. Some features in this work, most notably source analysis (*AN*), are geared towards a transfer-based system such as LMT. However, most features are architecture-independent, assume a "black-box" view of the system, and so are generically applicable for MT (White, 1995).

6 Conclusion

The fine-grained evaluation framework developed in this paper is a powerful new tool for MT system development. The detailed information it provides of translation quality, types and likely sources of translation errors, and their incidence in a given automatic translation task, is useful for MT system diagnosis and development planning, to better allocate development resources.

The proposed evaluation framework can be seen as one particular instantiation of a general framework, such as ISLE. The results of the evaluation experiments have confirmed the importance of continued development of the bilingual transfer lexicon and indicated the need for more accurate translation of named entities, which is addressed in an architectural improvement to the LMT system currently being implemented.

Acknowledgements

We thank thank C. Rodriguez and A. Meiss for help with the translation evaluations, and C. Gdaniec, M. McCord and W. Zadrozny for comments that helped improve the quality of the paper.

We also thank the conference evaluators for helpful review comments, and IBM Corporate Community Relations for its support of this work.

References

Arendse Bernth. 1999. A Confidence Index for Machine Translation. In *Proceedings of TMI-99*. University College, Chester, England.

LDC. 2002. Linguistic Data Annotation Specification. http://www.ldc.upenn.edu/projects/TIDES. Philadelphia, PA.

George Doddington. 2001. Automatic Evaluation of Machine Translation Quality using n-Gram Co-occurrence Statistics. NIST *http://www.nist.gov/speech/tests/mt*. Washington, DC.

EAGLES MT Evaluation Working Group. 1996. EAGLES Evaluation of Natural Language Processing Systems: Final Report. *EAGLES Document EAG-EWG-PR.2, ISBN 87-90708-00-8*. Center for Sprogteknologi, Copenhagen.

Claudia Gdaniec. 1999. Using MT for the Purpose of Information Assimilation from the Web. In *Workshop on Problems and Potential of English-German MT systems*. TMI, Chester, UK.

Edward Hovy. 1999. Toward Finely Differentiated Evaluation Metrics for Machine Translation. In *Proceedings of the EAGLES Workshop on Standards and Evaluation*. Pisa, Italy.

Edward Hovy, Margaret King, and Andrei Popescu-Belis. 2002. Computer-Aided Specification of Quality Models for Machine Translation Evaluation.

International Standards for Language Engineering (ISLE) 2000. The ISLE Classification of Machine Translation Evaluations. *http://www.isi.edu/natural-language/mteval*. USC Information Sciences Institute, Marina del Rey, CA

Michael McCord. 1989. Design of LMT: A Prolog-Based Machine Translation System. *Computational Linguistics*, Vol. 15, No. 1. MIT Press, Cambridge, MA.

Michael McCord and Arendse Bernth. 1998. Design of LMT: The LMT Transformational System. In *Proceedings of AMTA-98*. Association for Machine Translation in the Americas.

Sonja Nießen, Franz J. Och, Gregor Leuch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of LREC-2000*. Athens, Greece.

Eric Nyberg, Teruko Mitamura, and Jaime Carbonell. 1994. Evaluation Metrics for Knowledge-Based Machine Translation. In *Proceedings of COLING-1994*. Kyoto, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the ACL-2002*. Philadelphia, PA.

Florence Reeder, Keith Miller, Jennifer Doyon, and John White. 2001. The Naming of Things and the Confusion of Tongues: An MT Metric. In *MT Evaluation Workshop, MT Summit 2001*. Santiago de Compostella, Spain.

Florence Reeder. 2001. In One Hundred Words or Less. In *MT Evaluation Workshop, MT Summit VIII, 2001*. Santiago de Compostella, Spain.

Michelle Vanni and Keith Miller. 2001. Scaling the ISLE Framework: Validating Tests of Machine Translation Quality for Multi-Dimensional Measurement. In *MT Evaluation Workshop, MT Summit VIII, 2001*. Santiago de Compostella, Spain.

John White. 1995. Approaches to Black-box Machine Translation Evaluation. In *Proceedings of the MT Summit 1995*. Luxembourg.

John White. 2001. Predicting Intelligibility from Fidelity in MT Evaluation. In *MT Evaluation Workshop, MT Summit VIII, 2001*. Santiago de Compostella, Spain.

Yorick Wilks. 1994. Traditions in the Evaluation of MT. In *MT Evaluation: Basis for Future Directions*, M. Vasconcellos, ed. Proceedings of NSF workshop, San Diego, CA.

Appendix: LMT translation quality scale

Score definitions and criteria for "Intelligibility of Target Text" and "Fidelity to Source Text." [8]

Score	Criteria
5	The translated sentence is perfectly intelligible and correctly reflects the meaning of the original.
4	There are one or two minor syntactical, lexical or grammatical mistakes in the translation, but the target sentence is mostly intelligible and mostly reflects the meaning of the source.
3	There are one or two major or three or four minor syntactical, lexical or grammatical mistakes in the translation compromising the intelligibility of target and fidelity to source.
2	Most of the translation is not intelligible and/or true to the original.
1	The translated sentence is completely garbled and not intelligible at all.
0	The source sentence was not translated OR the translation is mostly intelligible, but in large part does not reflect the meaning of the original (severe mistranslation).

Table 6: LMT Translation Quality Scale

[8] The scale and evaluation criteria were created by Marga Taylor for a previous LMT evaluation.