

Comparative Evaluation of Online Machine Translation Systems with Legal Texts*

Chunyu Kit** and Tak Ming Wong***

The authors discuss both the proper use of available online machine translation (MT) technologies for law library users and their comparative evaluation of the performance of a number of representative online MT systems in translating legal texts from various languages into English. They evaluated a large-scale corpus of legal texts by means of BLEU/NIST scoring, a de facto standard way of exercising translation-quality evaluation in the field of MT in recent years and a method that provides an objective view of the suitability of these systems for legal translation in different language pairs.

¶1 Translation is one of those services for which the demand can never be satisfied. Machine translation (MT), the technology for the automation of translation, has long been hoped to be the answer. In this article, we will examine a number of the most commonly used online MT services and compare their translation performance on legal texts, a text genus of particular importance to law librarians and law library users.

¶2 MT evaluation has garnered attention during the more than forty years since the first ALPAC report.¹ Recently, Yates performed a user-oriented MT evaluation on legal text with twenty sentences, ten in Spanish and ten in German.² In her article, only one online MT system, Babel Fish,³ is evaluated from the user's perspective to determine whether it achieves the objective as stated on the site, that is, "to grasp the general intent of the original."⁴ Based on her manual evaluation of the Babel Fish translations of the twenty sentences by comparing them to the reference translation of them, Yates arrived at the conclusion that "Babel Fish is of little use in a law library."⁵

* © Chunyu Kit and Tak Ming Wong, 2008.

** Assistant Professor, Department of Chinese, Translation and Linguistics, City University of Hong Kong, Hong Kong SAR.

*** Research Student, Department of Chinese, Translation and Linguistics, City University of Hong Kong, Hong Kong SAR.

1. AUTOMATIC LANGUAGE PROCESSING ADVISORY COMMITTEE, LANGUAGE AND MACHINES: COMPUTERS IN TRANSLATION AND LINGUISTICS (1966) (Publication 1416, National Academy of Sciences National Research Council), available at <http://www.nap.edu/openbook.php?isbn=ARC000005>.
2. Sarah Yates, *Scaling the Tower of Babel Fish: An Analysis of the Machine Translation of Legal Information*, 98 LAW LIBR. J. 481, 491, 2006 LAW LIBR. J. 27, ¶ 42.
3. AltaVista, Babel Fish Translation, <http://babelfish.altavista.com> (last visited Dec. 21, 2007).
4. AltaVista, Babel Fish FAQ, Why Are Some Translations Lesser Quality?, http://www.altavista.com/help/babelfish/babel_faq (last visited Jan. 4, 2008).
5. Yates, *supra* note 2, at 500, ¶ 67.

¶3 In a reply to Yates, Somers pointed out many problems in her evaluation methodology and proposed an alternative evaluation scheme where a subject is asked a series of factual questions about a case based on a Babelfish translation of its case summary. According to Somers, “[t]his is not dissimilar to the use of reading comprehension tests, a technique that has been proposed quite frequently over the years” for the comparative evaluation of MT systems.⁶ He opines that because “[s]ummarizing the important aspects of a reported case is a genuine task for trainee and practicing lawyers . . . [i]n this way we would have an objective rather than subjective assessment of the fitness for task of Babelfish.”⁷ However, we do not know what kind of methodological problems this scheme is likely to have once it is put in practical use, e.g., how much of the evaluation outcome would come from a reader’s comprehension ability and how much from the quality of a translation, as Somers has not yet conducted this evaluation.⁸

¶4 This article, however, is not aimed at evaluating the translation quality and capacity of any particular MT system or at discussing the pros and cons of any particular evaluation methodology. Instead, we report on a horizontal comparison of different online MT systems carried out with a large volume of legal texts. Unlike the approach of evaluating MT quality by human judgment, which is in sharp contrast to the ordinary practice nowadays in the field of MT, our approach adopts the state-of-the-art automatic quantitative MT evaluation technology that has been commonly accepted and widely applied by MT developers and researchers in recent years.⁹ MT evaluation with larger data sets is undoubtedly more reliable in offering an objective instead of a subjective understanding of the performance of the MT systems in question, no matter which language pairs are involved.

¶5 The purpose of this article is to discuss the proper use in legal translation of the currently available MT technologies. Based on our quantitative evaluation using a large corpus of legal texts, we provide substantial evidence to assist in the proper selection of online MT systems for different language pairs, which is the very first step toward the proper use of them. Misunderstanding of MT can be avoided if it is put in a proper place in the industry of translation. Misconceptions and over-expectations about a particular MT system can be avoided if its performance on a particular genre of texts, e.g., legal texts, is properly and concretely evaluated with real data. Specifically, online MT can be utilized to facilitate legal translation to a great extent if necessary information is available to its users about

6. Harold Somers, *The Use of Machine Translation by Law Librarians—A Reply to Yates*, 99 LAW LIBR. J. 611, 618, 2007 LAW LIBR. J. 35, ¶ 34 (citing Harold L. Somers & Nuria Prieto-Alvarez, *Multiple-Choice Reading Comprehension Tests for Comparative Evaluation of MT Systems*, in WORKSHOP ON MACHINE TRANSLATION EVALUATION AMTA-2000, at 107 (2000)).

7. *Id.* at 618–19, ¶ 35.

8. *See id.* at 619, ¶ 35.

9. *See id.* at 616, ¶ 23.

their best choice for selecting an MT system to meet their translation demand in a particular language pair.

¶6 In the following sections, we first give a brief introduction to MT, providing a scenario for our work. Then, we present our quantitative evaluation of a number of representative online MT systems, including evaluation data, criteria, procedure, and results. We hope that our findings and discussions can lead to a deeper understanding of MT and hence enable our readers to get the maximum benefit from available online MT systems by using them for the translation tasks for which they are most suitable.

Proper Understanding of Machine Translation

¶7 The novelty of online MT services may give the impression that MT is something quite new. As a matter of fact, the MT research that has enabled such services had been ongoing for nearly half a century¹⁰ by the time Babel Fish was launched in late 1997.¹¹ Systran, the technology provider of Babel Fish, launched its first MT system in 1976 for the Commission of European Communities.¹² In the next three decades, however, according to John Hutchins's comparison of the translation outputs by MT systems in the 1960s, 1970s, and 1980s, MT technology failed to show substantial improvement in translation quality, a situation that may continue for the foreseeable future.¹³ His finding reminds us to reconsider another question that was brought to our attention by Martin Kay decades ago: what are the proper roles for MT and human translators in translation?¹⁴

¶8 From the developer's perspective, an article by Systran developers reports that:

[T]he majority of users are amazingly understanding of MT capabilities and limitations. They are impressed by the capability and even the translation quality.

....

Many users show that they know what to expect from MT: less than perfect results. They are willing to help.

-
10. Readers interested in MT history may refer to John Hutchins, *The History of Machine Translation in a Nutshell* (Nov. 2005), <http://www.hutchinsweb.me.uk/Nutshell-2005.pdf>.
 11. Kurt Ament, *Babel Fish*, <http://www.infotektur.com/demos/babelfish/en.html> (last visited Jan. 4, 2008).
 12. John Hutchins, *Machine Translation: A Concise History*, in *COMPUTER AIDED TRANSLATION: THEORY AND PRACTICE* (forthcoming 2008), available at <http://www.hutchinsweb.me.uk/CUHK-2006.pdf>, at 7 (last visited Jan. 4, 2008).
 13. John Hutchins, *Has Machine Translation Improved? Some Historical Comparisons*, in *IAMT, MT SUMMIT IX 7-8* (2003), <http://www.amtaweb.org/summit/MTSummit/FinalPapers/12-Hutchins-final.pdf>.
 14. Martin Kay, *The Proper Place of Men and Machines in Language Translation*, 12 *MACHINE TRANSLATION* 3 (1997).

....

Users also realize the challenges of MT: name handling, idiomatic expressions and context-sensitive translations.¹⁵

The user feedback for Systran also shows that users are not blind to the limitations and constraints of MT. Users with experience in using MT systems can discover what MT can and cannot handle and what MT is good and not good at handling.

Why Is It So Difficult for Machines to Translate?

¶9 In Yates's paper, the difficulties of MT were reviewed from the perspectives of the complexity of human language and translation.¹⁶ Briefly put, language is full of exceptions and ambiguities at all linguistic levels. While humans can recognize these extraordinary linguistic features and handle them properly, machines are incapable of performing the same job without adequate human intelligence. Current MT systems are said to focus only on linguistic analysis at morphological and lexical levels, and provide limited syntactical analysis, far from sufficient to support sophisticated translation, not to mention any kind of cultural understanding.¹⁷

¶10 We may look into this issue from the perspective of MT architecture. The well-known MT pyramid presented in figure 1 illustrates three classical MT approaches.¹⁸ The direct approach at the bottom and the interlingual approach at the top are beyond the scope of our discussion, for all MT systems we intend to examine take the transfer approach at the middle,¹⁹ although some (e.g., Google Translate) are more synthesis-heavy than the others.

15. Jin Yang & Elke Lange, *Going Live on the Internet*, in *COMPUTERS AND TRANSLATION* 191, 198–99 (Harold Somers ed., 2003).

16. See generally Yates, *supra* note 2.

17. See Yates, *supra* note 2, at 484, ¶ 14.

18. This MT pyramid diagram (with some modifications) was first used by Vauquois in 1968, according to Doug Arnold, *Why Translation is Difficult for Computers*, in *COMPUTERS AND TRANSLATION*, *supra* note 15, at 119, 122.

19. The transfer approach is also known as the rule-based approach.

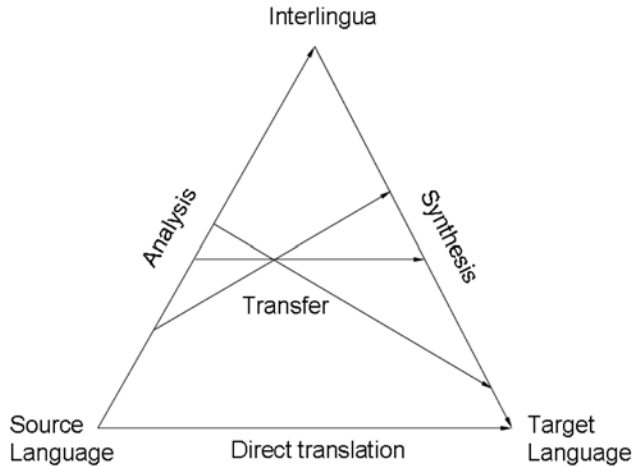


Fig. 1. MT Pyramid

¶11 Basically, transfer-based MT involves the following three main tasks:

1. *Analysis*, to transform the surface form of a source sentence into an abstract representation with respect to the linguistic characteristics of the source language.
2. *Transfer*; to map this representation for the source language to that for the target language.²⁰
3. *Synthesis*, to transform the abstract representation for a target sentence to its surface form.²¹

In the analysis stage, an input with ill-formed sentences needs to be handled properly. Unfortunately, however, most current transfer-based MT systems are not equipped with appropriate rules for this purpose. Also, ambiguities in natural language may present many problems unsolvable by machine, hindering it from selecting the right meaning, usually among several possible ones, that a sentence or phrase is intended to present.

¶12 There are at least two types of ambiguity commonly involved in translation. Lexical ambiguity concerns polysemous words that have more than one meaning and, therefore, more than one possible translation in another language. For example, the word “post” in “the post has arrived” may refer to “delivery of mail” or “a piece of wood.” Structural ambiguity concerns sentences with more than one possible structure. For example, it is unclear whether “yesterday” in “the minister

20. As shown in figure 1, there is no single way of transfer; some MT systems may adopt a synthesis-light approach (e.g., Systran) while some may adopt a synthesis-heavy approach (e.g., Google).

21. Arnold, *supra* note 18, at 122.

stated that the proposal was rejected yesterday” refers to the time of the proposal being stated or being rejected.²²

¶13 The complexity of the transfer process varies according to the linguistic diversities between the source and target languages. As different languages may express the same content with words and structures of some subtle difference, adequate rules must be formulated to bridge these diversities or gaps, among which the lexical gap is well known in translation. The following example illustrates a critical distinction between English and Spanish:

English: Apples are sold here.

Spanish: *Se venden manzanas aquí.*

Self they-sell apples here.

Lit.: “Apples sell themselves here.”²³

For such a simple expression, English uses a passive construction while Spanish uses a reflexive one. To translate them by machine, complex transfer rules are required.

¶14 At first glance the synthesis looks like the reverse process of the analysis. This is conceptually true. In practice, however, analysis and synthesis deal with totally different issues. The former determines an intermediate representation for the meaning of an input sentence, whereas the latter selects a surface sentence form for an intermediate representation of the same meaning. Since there can be more than one grammatical way to express the same meaning, it is very difficult for a machine to decide the proper one among so many possibilities. For example:

- a. What time is it?
- b. How late is it?
- c. What is the hour?²⁴

Only (a) is correct, but there is hardly any reason to explain why it is right and the others are not in a particular language. Most problems concerning sentence synthesis are not always about right or wrong. More often, there seems no reason to select a particular one among several alternatives. For example:

- a. Sam saw a black cat.
- b. Sam saw a cat. It was black.
- c. Sam saw something black. It was a cat.
- d. Sam saw a cat which was black.
- e. A black cat was seen by Sam.²⁵

22. *Id.* at 124–25.

23. *Id.* at 129.

24. *Id.* at 133–34.

25. *Id.* at 133.

The same content can be expressed by (a)–(e), and possibly many others. Without contextual information, no one can determine which one is better than another. This kind of decision making seems beyond the machine’s capability for the time being.

¶15 One should note as well that translation is difficult not just for machines, but even for humans. Translation is beyond human translators’ reach if they are not equipped with a proper knowledge of the subject of the text to be translated. Learning to improve by themselves is critical to translators, no matter whether they are machines or humans. Otherwise, they will repeat the same errors endlessly.²⁶

Proper Use of MT in Legal Translation

¶16 The quality of MT output is constrained, to a great extent, by that of the source text input. A text with proper grammar and unambiguous wording often leads to “gistable” translation by machine, allowing readers to understand the general meaning of the source text and determine its relevance to their research.²⁷ However, as the developers of Babel Fish note, a text with “slang, misspelled words, poorly placed punctuation, and complex or lengthy sentences” can cause the text to be translated incorrectly.²⁸

¶17 It is true that legal translation involves many of the above-mentioned problems, e.g., extraordinarily long sentences, and, in consequence, produces unsatisfactory MT outputs. Yates has discussed the difficulties of legal translation in general.²⁹ Briefly, the lengthy and complex sentences with unusual word order, plus the use of culture- and legal system-dependent terminology, have made legal texts, in a sense, a special type of language. The lack of one-to-one correspondence between legal concepts also puts a damper on any attempt to bridge the divide between two legal systems.³⁰ Current MT systems are considered unsuitable for legal translation if not equipped with adequate legal and cultural knowledge. “It is true now, and will probably always be true,” commented Hutchins.³¹ From this perspective, the usefulness of a machine-translated legal text is, of course, always questionable.³²

26. See John Hutchins, *Current Commercial Machine Translation Systems and Computer-Based Translation Tools: System Types and Their Uses*, 17 INTERNATIONAL JOURNAL OF TRANSLATION 5, 34 (2005) (discussing reasons why MT fails and how it can be improved using feedback).

27. See Free Translation.com Help and FAQ, What Should Machine Translation Be Used For?, <http://www.freetranslation.com/help/#uses> (last visited Jan. 10, 2008).

28. AltaVista Help, Translation, http://www.altavista.com/help/babelfish/babel_help (last visited Jan. 10, 2008).

29. Yates, *supra* note 2, at 486–88, ¶¶ 21–29.

30. *Id.* at 486, ¶ 25.

31. John Hutchins, *Machine Translation and Computer-based Translation Tools: What’s Available and How It’s Used*, in A NEW SPECTRUM OF TRANSLATION STUDIES 13, 42–43 (José Maria Bravo ed., 2004).

32. Terence Lewis, *When Not to Use MT and Other Translation Tools* 36–38 (1997), <http://www.mt-archive.info/EAMT-1997-Lewis.pdf> (paper presented at EAMT Workshop, Copenhagen, May 1997).

¶18 Instead of further blaming MT for its uselessness, however, we suggest that MT can be a good solution in certain situations, even in the context of legal translation. Clearly, MT is incapable of translating “any kind of text in any subject” and producing “unaided a good translation.”³³ It is not designed, nor proposed, to do so. When considering the usability of MT, we have to bear in mind that usability is relative to users’ expectations. In a survey exploring the usefulness of MT from the users’ perspective, it was noted that “those who feel comfortable with English do not want pure MT translations, but those who are not as strong in English find it useful.”³⁴ Thus, a fair comment from this survey is that “Pure MT is rough—often obscure, frequently humorous—but it can be useful.”³⁵ This suggests a direct benefit that MT can provide: because human translators are not always available, MT can many times serve as a “good-enough” solution.³⁶

¶19 Indeed, a better view of MT as a usable tool requires a sound understanding of the demands that translation is expected to answer. According to Hutchins, the demands can be categorized into four types (described in table 1) according to their purposes and the quality of translation required to serve such purposes.³⁷

Table 1

Four types of translation demand and their requirements on translation quality

Purpose of translation	Required quality of translation
Dissemination	Publishable quality
Assimilation	At a lower level of quality
Interchange	Translation between participants in one-to-one communication or of an unscripted presentation
Information Access	Translation within multilingual systems of information retrieval, information extraction, database access, etc.

¶20 Undoubtedly, professional legal translation belongs to the first type. As discussed above, currently MT cannot be relied upon to produce legal translation of publishable quality. However, it is useful for other purposes. In general, MT can serve quite well when translation quality is not the first priority, e.g., when what a

33. Hutchins, *supra* note 26, at 34.

34. D. Verne Morland, *Nutzlos, Bien Pratique, or Muy Util? Business Users Speak Out on the Value of Pure Machine Translation*, <http://www.roi-learning.com/dvm/pubs/articles/tatc-24> (last visited Mar. 18, 2008).

35. *Id.*

36. MIKE DILLINGER & ARLE LOMMEL, LISA BEST PRACTICE GUIDES: IMPLEMENTING MACHINE TRANSLATION 4 (2004), available at <http://www.lisa.org/products/bestPractice/index.html> (site registration required).

37. See John Hutchins, *The Development and Use of Machine Translation Systems and Computer-based Translation Tools*, 15 INTERNATIONAL JOURNAL OF TRANSLATION 5–6 (2003).

user needs to know is the rough idea or the subject of a text, so as to locate information or “decide whether or not to have a human translator to provide a publication-quality translation.”³⁸

¶21 Currently there are a large variety of online MT systems to provide nearly instant translation in almost all domains, far faster than human translators can, and free of charge. They serve all popular languages. Their translation quality varies from system to system, from language to language, and from text to text. Our interest here is to examine their quality in the domain of legal translation. Our comparative evaluation will show how well they translate legal texts for different language pairs. By submitting the test texts to the target MT systems and comparing the quality of their translation outputs, we can arrive at a sound understanding of the usefulness of the currently available online MT technologies in legal translation and come up with a clear idea of which online MT system is the best choice for automatic translation of legal texts for a particular language pair.

Quantitative MT Evaluation

¶22 It is a difficult task to evaluate translation, including machine translation. As noted by White, there is no ideal or correct translation for a text among so many possible ways to translate it.³⁹ But without a benchmark to compare with, there is hardly any trustworthy quantitative measure that we can follow to quantify the quality of MT output. If we resort to manual evaluation, other problems arise, including human judges’ personal bias and the inevitable inconsistency of their judgments. Many human factors can lead to unreliable evaluation, among which the following three are particularly worth noting:

1. *History*: Many things irrelevant to the evaluation can intervene in human judgment. For example, the result of a World Cup final, a stock-market crash, one’s mood, and even weather may significantly influence the reliability of a human evaluator’s work, although training and experience may alleviate such negative impact to a great extent.
2. *Testing*: Human evaluators have a different sensation, and reaction, to things (e.g., an expression) between the first time and the second time they see them. The second time, they have an informed idea of what the expression is supposed to say. The more informed, the more easily they read. This may affect their judgment about the quality of the translation.
3. *Maturation*: Many ordinary situations can affect people’s ability to be consistent in their judgments. They can give a significantly different grade to the same translation sentence if they are tired, bored, hungry, unhappy, or fed up with the evaluation process.⁴⁰

38. DILLINGER & LOMMEL, *supra* note 36, at 8.

39. John S. White, *How to Evaluate Machine Translation*, in *COMPUTERS AND TRANSLATION*, *supra* note 15, at 211, 213.

40. *Id.* at 218–19.

In order to minimize the side effects of these and other human factors in MT evaluation, it is necessary to resort to automatic evaluation with a large data set so as to obtain a more objective, and thus more reliable, evaluation result. Manual evaluation with large data sets can be very time-consuming and costly, especially when a large set of texts and a large number of language pairs are involved for more than one MT system.

¶23 A well-known approach to MT evaluation called “round-trip translation” or “back-and-forth translation” has been used frequently by the public, from journalists to lay users, whereby a text is translated into the target language and then back by the same MT system.⁴¹ Although at first glance it looks fair, this technique is rather questionable.

¶24 Most of its supporters favor it, thinking, “in theory, the back translated English should match the original English.”⁴² However, the dangers of this approach have been explained from the perspective of MT researchers and developers. The following advice of O’Connell concerning round-trip translation is quoted by Somers:

A common misunderstanding about MT evaluation is the belief that back translation can disclose a system’s usability. [. . .] The theory is that if back translation returns [the source language] input exactly, the system performs well for this language pair. In reality, evaluators cannot tell if errors occurred during the passage to [the target language] or during the return passage to [the source language]. In addition, any errors that occur in the first translation [. . .] cause more problems in the back translation.⁴³

A more straightforward explanation for the poor reliability of the round-trip translation can be found on the help page of an online MT system web site:

When you translate the text into another language, some words can be translated in a number of ways. . . . Consequently when the text is translated back into the original language these words may be translated differently.⁴⁴

According to Somers, this round-trip approach for MT evaluation is not a good way to identify which MT system is better or to tell how well a text is translated.⁴⁵

-
41. Harold Somers, *Round-Trip Translation: What Is It Good For?*, PROC. AUSTRALASIAN LANGUAGE TECH. WORKSHOP 127, 127 (2005), available at <http://www.alta.asn.au/events/altw2005/cdrom/pdf/ALTA200519.pdf>.
 42. Biomedical Translation, *More Machine Translation: Fun with Computer Generated Translation!* (Oct. 2003), <http://www.biomedical.com/news.html>.
 43. Somers, *supra* note 41, at 127–28 (quoting Theresa A. O’Connell, *Preparing Your Website for Machine Translation: How to Avoid Losing (or Gaining) Something in the Translation* (2001), <http://www-128.ibm.com/developerworks/web/library/us-mt>) (alterations in original).
 44. Free Translation.com Help and FAQ, *Why Does the Translation Not Make Sense If I Translate My Text into Another Language and Then Back Again?*, <http://www.freetranslation.com/help/#sense> (last visited Jan. 9, 2008).
 45. See Somers, *supra* note 41, at 131.

¶25 Therefore, we have to turn to some more serious, objective, and reasonable qualitative evaluation for MT technologies. For this purpose, Papineni et al. proposed the BiLingual Evaluation Understudy (BLEU) method,⁴⁶ with a language-independent statistical metric to provide a quick overview about the performance of an MT system by comparing its translation against available reference translation by a human translator. The rationale for measuring the quality of MT output this way follows a very simple hypothesis: “The closer a machine translation is to a professional human translation, the better it is.”⁴⁷ It has also been explained this way:

[E]ven though there are many different ways to translate a text “correctly”, most of them will share certain phrases in common. If this is correct, then it should be possible to model statistically a quasi ideal of that text against which translations can be compared in relatively simple string-by-string matches.⁴⁸

¶26 Practically, then, what BLEU evaluation needs is one or more professional human translations of a source text to serve as the gold standard to compare against its machine translation output. The evaluation then calculates the closeness between the MT output and the human translation by a numerical metric, namely, the BLEU metric. Its basic calculation, to put it in simplified terms, is to count the number of common words and word sequences (technically referred to as *n*-grams) shared by a translation candidate and a reference translation. This idea is illustrated in Papineni et al. with the following example:

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

...

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.⁴⁹

Candidate 1 shares twelve words with the reference translation, while candidate 2 shares only seven. For a fair comparison, these numbers are normalized by the lengths of the candidates, giving 12/18 and 7/14 (i.e., 0.67 and 0.5) for Candidate 1 and Candidate 2 respectively, indicating that the first candidate is preferable to the second. This example shows the basic idea underlying the BLEU metric.⁵⁰

46. See generally, Papineni, et al., *BLEU: A Method for Automatic Evaluation of Machine Translation*, in PROCEEDINGS OF THE 40TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 311. But see Somers, *supra* note 7, at 616, ¶ 23 for a critique of this method.

47. Papineni et al., *supra* note 46, at 311 (emphasis omitted).

48. White, *supra* note 39, at 240.

49. Papineni et al., *supra* note 46, at 312.

50. The details of the BLEU metric, which involves some more complex mathematical calculation such as *n*-gram precision and sentence brevity penalty, are not described here.

¶27 Following the development of the BLEU metric, the National Institute of Standards and Technology (NIST) developed another scoring facility based on BLEU, known as the NIST metric.⁵¹ The main difference between the two is that BLEU focuses more on language fluency while NIST focuses more on lexical accuracy. Interestingly, it is a widely accepted practice to use both metrics for a wider and more balanced view of the two essential characteristics of translation.

¶28 The BLEU and NIST metrics have been widely used by MT developers to monitor the improvement of their systems under development. In 2005, NIST began performing a yearly MT evaluation, using BLEU and NIST metrics as standard measurement, with large-scale data sets for the state-of-the-art MT systems.⁵² The NIST 2006 Machine Translation Evaluation (MT-06) involved forty-six MT systems and two language pairs, with a test set of 80,000 words for each language pair, summing up to 736,000 words in total.⁵³ Manual evaluation on such a scale would cost a huge amount of human effort, time, and money, and is thus very unlikely to take place.

¶29 By virtue of the reliability of computerized evaluation of this kind, a series of tests to compare the BLEU and NIST scores against human judgment of translation quality have been conducted. The results confirm that, as far as translation adequacy and fluency are concerned, both evaluation metrics correlate with human assessments sensitively and consistently.⁵⁴ Although automatic evaluation metrics are not supposed to be used to replace human judgment completely, they have demonstrated their great value in enabling large-scale MT evaluations at a controllable cost.

Methodology

Evaluation Data

¶30 For a fair comparison of the performance of different online MT systems on legal translation, we needed an appropriate set of legal texts in the languages that the MT systems to be evaluated are designed for. The texts selected for this

-
51. George Doddington, *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*, 2 HUMAN LANGUAGE TECHNOLOGY CONFERENCE 138, 138 (2002), available at http://portal.acm.org/ft_gateway.cfm?id=1289273&type=pdf&coll=GUIDE&dl=GUIDE&FID=11617123&CFID=54214794 (site registration required).
 52. NIST 2005 Machine Translation Evaluation Official Results (Aug. 1, 2005), http://www.nist.gov/speech/tests/mt/2005/doc/mt05eval_official_results_release_20050801_v3.html.
 53. See NIST, *The 2006 NIST Machine Translation Evaluation Plan* (Mar. 16, 2006), http://www.nist.gov/speech/tests/mt/2006/doc/mt06_evalplan.v4.pdf.
 54. Doddington, *supra* note 51, at 142, fig. 6.

evaluation included four European treaties⁵⁵ acquired from EUR-Lex,⁵⁶ a web site providing free access to many European Union law documents; and an official document from the web site of the United Nations, the Universal Declaration of Human Rights.⁵⁷ All of them are multilingual texts of a very high standard in each language, translated by professionals into most European and many other languages. These documents provided a total of thirteen language versions⁵⁸ that are covered by the online MT systems involved in this evaluation. The entire corpus of the five documents consists of 141,000 words for its English version, larger than the above-mentioned NIST MT Evaluation test set. It is therefore a reliable test set for examining the translation quality of legal texts by MT systems and to produce substantial statistical analysis for their performance in quantitative evaluation.

MT Systems Chosen for Evaluation

¶31 Six popular online MT systems were selected for this evaluation: Babel Fish,⁵⁹ Google,⁶⁰ ProMT,⁶¹ SDL free translator,⁶² Systran,⁶³ and WorldLingo.⁶⁴ It is worth noting that the first two are both powered by Systran's MT technology,⁶⁵ although Google has developed its own MT engines for two language pairs—Arabic to English and Chinese to English.⁶⁶ All of these systems can be freely accessed via the web and provide a usable online MT service. Usefulness was a primary criterion in selecting online MT systems for the evaluation. We required all systems to be able to handle a long sentence of at least one hundred words. Long sentences are very common in legal texts, and it is undesirable to cope with the word limit by segmenting a long source sentence into fragments for translation, for doing so may alter its meaning. This criterion eliminated those online MT systems with too strict a constraint on word limit.⁶⁷

-
55. Treaty of Nice, Mar. 10, 2001, 2001 O.J. (C 80) 1; Consolidated Version of the Treaty Establishing the European Community, Nov. 10, 1997, 1997 O.J. (C 340) 3; Consolidated Text of the Treaty on European Union, Dec. 24, 2002, 2002 O.J. (C 325) 5; Treaty Establishing a Constitution for Europe, Oct. 29, 2004, 2004 O.J. (C 310) 1.
 56. EUR-Lex, <http://eur-lex.europa.eu/en/index.htm> (last visited Jan. 4, 2008).
 57. The Universal Declaration of Human Rights, <http://www.unhchr.ch/udhr/index.htm> (last visited Jan. 4, 2008).
 58. The thirteen languages are Dutch, French, German, Greek, Italian, Portuguese, Russian, Spanish, Swedish, Arabic, Chinese, Japanese, and Korean.
 59. Babel Fish, *supra* note 3.
 60. Google Translate, http://www.google.com/language_tools (last visited Jan. 10, 2008).
 61. ProMT, <http://www.online-translator.com> (last visited Jan. 10, 2008).
 62. SDL free translator, <http://www.freetranslation.com> (last visited Jan. 10, 2008).
 63. Systran, <http://www.systransoft.com> (last visited Jan. 10, 2008).
 64. WorldLingo, http://www.worldlingo.com/en/products_services/worldlingo_translator.html (last visited Jan. 10, 2008).
 65. See Mary Flanagan & Steve McClure, *SYSTRAN and the Reinvention of MT 3* (Jan. 2002), <http://www.systransoft.com/DocDownload/industry-reports/2002.01.SYSTRAN.And.The.Reinvention.Of.MT.pdf>.
 66. See Posting of Franz Och to Google Research Blog, <http://googleresearch.blogspot.com/2006/04/statistical-machine-translation-live.html> (Apr. 28, 2006).
 67. For example, Reverso (www.reverso.net/text_translation.asp) has a limit of 400 characters for each translation, and Amikai (www.amikai.com/demo.jsp) a limit of 100 characters.

Issues Examined

¶32 Our study was not limited to determining whether the current MT technology is capable of translating legal texts at a high quality level or is suitable for legal translation at all. Legal translation is complex and probably will not be able to be handled reliably by machines in the foreseeable future.⁶⁸ Instead, our purpose here was to find out which of the currently available online MT systems can be a proper, or even the best, choice for translating legal texts between different languages, assuming an actual need for the practical use of MT for legal translation. In this evaluation we examined three critical issues concerning online MT service for legal translation: language coverage, translation quality for a particular language, and translation quality for a particular language pair, with a focus on the last.

¶33 *Language coverage:* The dominant language pairs used by law librarians are those from a foreign language to English.⁶⁹ Thus, the very first task in our evaluation was to look into the “to-English” language pairs served by each MT system. Also, the more language pairs an MT system serves, the less likely it is that its users will have to switch to another MT system, although such a switch may be inevitable if a user insists on having the best online MT service for each language pair.

¶34 *Translation quality for a language:* The complexity of translation lies largely in the linguistic diversity between languages. Intuitively, a larger gap between two languages would give rise to more difficulties in bridging them by translation, and consequently translation quality is more likely to be problematic. This is true for all MT systems. The performance of various systems on different language pairs in terms of the BLEU/NIST scores can serve as a faithful indicator of how well a particular language can be translated and how confident MT users can be with them.⁷⁰

¶35 *Translation quality for a language pair:* Within an MT system, the translation modules for different language pairs are actually, in a certain sense, different micro-MT systems. They do not necessarily all work in exactly the same way or with the same technologies, and may not all have the same scale of language resources. Their performance variance can reveal the integrated effect of such differences on different language pairs. Our evaluation in this situation is conducted in the hope that its results can help readers to select a proper MT system for a particular language pair.

Evaluation Procedure

¶36 We first visited the web sites of the candidate MT systems and collected the information about the language pairs they cover. Then the evaluation texts of the

68. See *supra* ¶ 6.

69. Sarah Yates, *I Need This in English*, AALL SPECTRUM, Apr. 2005, at 8.

70. In MT developers' terminology, how well a text or a language pair can be translated is termed “MT Translatability.”

five legal documents in each of these languages were acquired by extracting the text from their web pages.⁷¹ The resulting texts in languages other than English were used as the source texts and their English counterparts as “gold standard” target texts for the evaluation. The evaluation was performed by comparing each candidate MT system’s translation outputs for the source texts against the “gold standard” target texts. We developed a few computer programs for submitting the source texts sentence by sentence to each MT system and then collecting their translation outputs in an automatic manner. The computation of BLEU/NIST score was then carried out for each translation output and its source text using the MT scoring software provided by the speech group of NIST.⁷²

Evaluation Results

Language Coverage

¶37 The to-English language pairs served by the six representative online MT systems selected for our evaluation are presented in table 2, where all languages on the left are source languages. Note that “Chinese (traditional) to English” is not officially listed as a language pair supported by Google, but our test showed that Google produces exactly the same translation outcomes for it as for “Chinese (Simplified) to English,”⁷³ so we considered it one of the language pairs served by Google.

¶38 Table 2 shows that among the six MT systems, WorldLingo offers more choices of language pairs than the others for to-English translation.⁷⁴ The runner-up is Systran, followed by Babel Fish and Google. Interestingly, Babel Fish and Systran use the same translation engine but serve slightly different language pairs. Babel Fish offers one language pair, Greek to English, not available from Systran.

-
71. The extraction of textual content from a web page was carried out by trimming off all HTML tags and other noncontent codes in the page.
 72. MT Scoring Software (mteval-v11b.pl) (released May 20, 2005), <http://www.nist.gov/speech/tests/mt/2008/scoring.html>.
 73. Traditional and Simplified Chinese are encoded by two different Chinese encoding schemes, namely, Big5 and GB codes respectively. In our test, we found that Google Translate is capable of selecting the proper encoding between the two automatically.
 74. In fact, WorldLingo stands out significantly for its language coverage over all language pairs, by not restricting to a small number of fixed language pairs. It provides separate lists of source and target languages in its Web interface to allow its users to select any combination of them. This offers a total of 182 language pairs, far exceeding any other online MT system. Worldlingo, *supra* note 64.

Table 2

*Translation Service for to-English Language Pairs Provided
by Representative Online MT Systems⁷⁵*

Languages	Online MT Systems						
	BabelFish	Google	PROMT	SDL	Systran	WorldLingo	
European Language	Dutch	*			*	*	*
	French	*	*	*	*	*	*
	German	*	*	*	*	*	*
	Greek	*					*
	Italian	*	*	*	*	*	*
	Portuguese	*	*	*		*	*
	Russian	*	*	*	*	*	*
	Spanish	*	*	*	*	*	*
	Swedish					*	*
Middle East Language	Arabic		*		*	*	
Asian Language	Chinese (Simplified)	*	*			*	*
	Chinese (Traditional)	*	*			*	*
	Japanese	*	*		*	*	*
	Korean	*	*	*	*	*	*

Translation Quality for Language Pairs

¶39 Figure 2 presents the BLEU/NIST scores for the overall performance of the selected MT systems on translating the test set from various languages into English. Each score shows, in a comparative way, how well the test set is translated from a particular language by a particular online MT system. Among the European languages, we can see that French and Dutch receive consistently better results than the others, in terms of both BLEU (focusing more on language fluency) and NIST (focusing more on lexical accuracy) scores. The worst are Greek and Swedish. Putting table 2 and figure 2 side by side, we can observe an interesting correlation: a language pair more popular across the MT systems receives, in

75. Data as of Jan. 13, 2007.

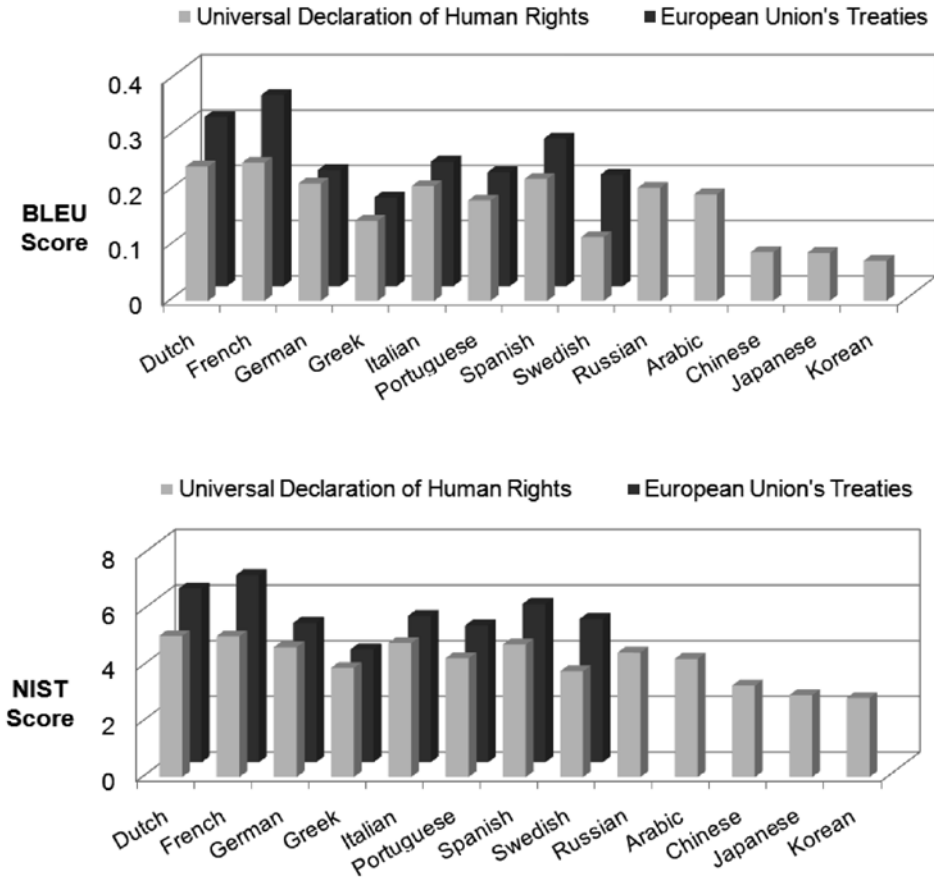


Fig. 2. The Average Performance of the Online MT Systems on Translation of the Test Set from Various Languages into English

general, a better online translation. A noticeable exception is Dutch. Its scores are highly comparable to those for French, although it is served by only three of the six MT systems. A general trend of this kind appears to be related to the market demand for translation between popular languages. A greater market demand certainly draws more investment, resources, and human effort, and hence results in a better service. On the contrary, Greek and Swedish, which are served by only two of the MT systems, are relatively poor. Even so, they receive much better scores than Asian languages.

¶40 The MT systems' performance on the three Asian languages: Chinese, Japanese, and Korean, involved in our evaluation, are remarkably poorer than that on other languages. Such a sharp contrast seems related to linguistic diversity between languages, in addition to market demand. In theory, the greater the linguistic diversity between two languages, the more difficult to translate between them. The three Asian languages differ greatly from English in terms of their idiosyn-

cratic linguistic features and structures at all morphologic, syntactic, and semantic levels. This may account for the poor performance of automatic translation from them into English. Interestingly, however, linguistic diversity seems not powerful enough to prevent Arabic from receiving better evaluation scores over five languages: three Asian and two European.

Translation Quality of MT Systems

¶41 The BLEU/NIST scores for each online MT system's performance on translating the test set from various languages to English are presented in tables 3a and 3b. The highest score for each language pair is underlined. Also, the scores within a 5% range of the highest score for each language pair are shaded. The resulting gray zones present the best online MT systems for the language pairs involved in our evaluation.

¶42 In general, all six MT systems' performance on translating European languages into English are relatively close to each other. This performance is overwhelmingly better than that for Asian languages. The performance for Arabic is generally positioned between these two groups, though it is better than that for two of the European languages. An exceptional case is Google's performance on translating Arabic to English, illustrating the highest scores in our evaluation. Another notable finding is that the MT systems powered by Systran technologies, i.e., Systran, Babel Fish, and Google, perform comparatively better than others. Interestingly, the latter two seem to have their translation engines further fine-tuned or supplemented with more linguistic resources, enabling them to achieve even better translation outcomes than Systran for many language pairs. Given that "a [BLEU] score of around 0.4 is usually considered to represent a fairly good translation,"⁷⁶ the translation of the European Union's Treaties by four of the six online MT systems, namely, Google, Babel Fish, Systran, and WorldLingo, has approached this quality level very closely, according to the results presented in table 3a.

¶43 Nevertheless, different systems show different strengths and weaknesses for different languages. Babel Fish is the best choice for many language pairs, especially for translation from Greek and Russian to English. Google outperforms its competitors outstandingly on translating Arabic and Chinese to English. These two are the official language pairs in the NIST 2005 and 2006 MT Evaluation.⁷⁷ They seem to have provided a real stage for Google to demonstrate the power of its MT engines. PROMT works slightly better than others on Portuguese, and WorldLingo appears to be a better choice than Systran for Swedish to English translation, although its quality is still low.

76. Somers, *supra* note 7, at 616 n. 29.

77. NIST, *supra* note 52; NIST, *supra* note 53, at 1.

Table 3a

The BLEU Scores for the Online MT Systems' Translation of the Test Set from Various Languages into English

Documents	Languages	Online MT Systems					
		BabelFish	Google	PROMT	SDL	Systran	WorldLingo
Universal Declaration of Human Rights	Dutch	<u>0.2576</u>			0.2051	0.2548	0.2538
	French	0.2620	<u>0.2666</u>	0.2451	0.2103	0.2616	0.2546
	German	0.2062	0.2448	0.1327	0.2067	<u>0.2470</u>	0.2363
	Greek	<u>0.1501</u>				0.1448	0.1392
	Italian	0.2151	0.2207	0.2144	0.1613	<u>0.2210</u>	0.2127
	Portuguese	<u>0.1881</u>	0.1880	0.1797	0.1604	<u>0.1881</u>	0.1853
	Russian	<u>0.2231</u>		0.2146		0.1820	0.1989
	Spanish	0.2212	0.2207	0.2184	<u>0.2270</u>	0.2207	0.2169
	Swedish					0.0844	<u>0.1461</u>
	Arabic		<u>0.5085</u>			0.0353	0.0346
	Chinese	0.0594	<u>0.1650</u>			0.0593	0.0686
	Japanese	<u>0.0888</u>	0.0804			<u>0.0888</u>	<u>0.0888</u>
	Korean	0.0747	0.0709			<u>0.0748</u>	0.0701
European Union's Treaties	Dutch	<u>0.3498</u>			0.2127	0.3382	0.3245
	French	0.3871	<u>0.3913</u>	0.2898	0.2470	0.3861	0.3739
	German	0.2395	<u>0.2400</u>	0.1491	0.1730	0.2341	0.2246
	Greek	<u>0.2099</u>					0.1104
	Italian	<u>0.2337</u>	0.2320	0.2330	0.2043	<u>0.2298</u>	0.2178
	Portuguese	0.2091	0.2082	<u>0.2182</u>	0.1981	0.2065	0.1961
	Spanish	0.2807	<u>0.2808</u>	0.2426	0.2640	0.2803	0.2532
	Swedish					0.1968	<u>0.2065</u>

Conclusion

¶44 We have tried in this article to construct a comparative view of the translation performance of representative online MT systems via empirical evaluation with a large-scale test set of real legal texts using BLEU/NIST metrics. BLEU/NIST scoring has been the *de facto* standard, and thus the most authentic and authoritative way of qualitative evaluation in the field of MT in recent years. This practical

Table 3b

The NIST Scores for the Online MT Systems' Translation of the Test Set from Various Languages into English

Documents	Languages	Online MT Systems					
		BabelFish	Google	PROMT	SDL	Systran	WorldLingo
Universal Declaration of Human Rights	Dutch	<u>5.2485</u>			4.5883	5.2154	5.2400
	French	5.2306	<u>5.2422</u>	4.8743	4.6238	5.2232	5.1673
	German	4.6484	5.0790	3.5692	4.5991	<u>5.1038</u>	5.0023
	Greek	<u>4.0209</u>				3.8597	3.8972
	Italian	4.9335	<u>4.9814</u>	4.8592	4.2266	4.9848	4.9162
	Portuguese	4.3062	4.3302	<u>4.3379</u>	4.0241	4.3062	4.3318
	Russian	<u>4.7452</u>		4.3726		4.2839	4.4872
	Spanish	4.8298	<u>4.8316</u>	4.5447	4.7691	<u>4.8316</u>	4.8002
	Swedish					3.3538	<u>4.2529</u>
	Arabic		<u>7.3365</u>			2.7348	2.6774
	Chinese	3.0591	<u>3.8392</u>			3.0579	3.1936
	Japanese	<u>2.9963</u>	2.7942			<u>2.9963</u>	2.9887
	Korean	2.8632	2.8212			<u>2.8662</u>	2.8068
European Union's Treaties	Dutch	<u>6.8559</u>			5.1465	6.6033	6.3832
	French	7.1564	<u>7.2089</u>	6.1052	5.7232	7.1355	7.0534
	German	<u>5.4683</u>	5.4301	3.9825	4.7778	5.2995	5.1189
	Greek	<u>5.2819</u>					2.8356
	Italian	<u>5.3962</u>	5.3834	5.3854	4.9232	5.3189	5.1586
	Portuguese	4.9577	4.9197	<u>5.0501</u>	4.9443	4.9002	4.7437
	Spanish	<u>5.9053</u>	5.8886	5.2738	5.6942	5.8867	5.5001
Swedish					5.0274	<u>5.2783</u>	

approach to MT evaluation incorporates human judgment by using human translation as a “gold standard” reference to compare against MT outputs. It relies on large-scale test data and, consequently, gives more reliable, objective, and consistent evaluation results than manual evaluation.

¶45 The usability and limitations of current MT technologies are analyzed from the perspectives of MT infrastructure and users' demands for translation quality. The proper use of currently available imperfect technologies to serve a

user's purpose well is a ubiquitous problem in many domains, including translation. While some MT developers have announced positive news about the advancements of MT technology, predicting or expecting MT to achieve human translation quality,⁷⁸ cautious MT researchers have pointed out that "over the last ten years there has been hardly any improvement in MT quality for many language pairs"⁷⁹ It is arguable whether it is realistic to expect MT to reach this level of translation quality in the foreseeable future. However, it is always preferable to have machines to take over as much routine work as possible for translation, such as dictionary look-up and the memorization and retrieval of existing translations, so as to free human translators for the more intelligent, artistic and creative parts of translation, which could and should never be taken over by machines.⁸⁰ Interestingly, along with the rapid technological advancements of MT, inferring an optimal choice of translation for a word, a phrase, or a sentence will become one of the routine tasks carried out by machine. Working on top of this, human translators are expected to concentrate more on improving the quality of MT outputs with their language intuition, knowledge, and creativity to achieve greater productivity in translation.

¶46 The focus of this study is more on the proper use of available MT service than on MT technology development. Given the various MT systems freely available for online translation service for different language pairs with different strengths and weaknesses, we have demonstrated a quantitative evaluation approach with the BLEU/NIST metrics as an objective means to compare their performance of legal translation on a large-scale corpus of legal texts. We hope that the evaluation results from a data set larger in volume than the NIST 2006 MT Evaluation test set can help MT users in the field opt for a proper MT system for their translation purpose. As a whole, we did not observe any particular MT system outperforming others for all language pairs. Different systems perform differently on different language pairs. Noticeably, however, in general the Systran family (including Babel Fish, Google, and Systran) prevailed over others.

¶47 Only one single text genus, namely, legal texts, was used in our evaluation, and the translation was restricted to the language pairs from a few popular languages to English. A more comprehensive evaluation with more text genera in more language pairs would undoubtedly give a more complete picture about the full potentials of the online MT systems. A manual qualitative evaluation⁸¹ on a substantial volume of data, if possible, would certainly help us to understand their translation performance and characteristics much better, or at least more intuitively. Also, as MT is a kind of computer application, program usability testing, as White

78. See Jan Krikke, *Machine Translation Inching Toward Human Quality*, IEEE INTELLIGENT SYSTEMS, Mar./Apr. 2006, at 4.

79. Hutchins, *supra* note 26, at 30.

80. See Kay, *supra* note 14, at 3, 20–22.

81. An evaluation similar to the one performed by Yates. See Yates, *supra* note 2.

proposed,⁸² is another route to show the power of MT systems. All these are beyond the scope of this study.

¶48 It is a fact that the MT quality is still far from satisfactory, let alone perfect, and the “gistable” translation by MT is far from publishable. However, it should also be recognized that the currently available MT technologies, if properly utilized, have been good enough to serve most of the translation demands for the purposes of information access, interchange, and assimilation. Even for the purpose of information dissemination, professional use of MT can minimize one’s workload towards publishable translation. In this regard, MT is more specifically aimed at enhancing human translators’ productivity and creativity by not only releasing them from routine work, which is best for machines, but also by providing them with more translation possibilities and, more importantly, the best choices for translating each word, phrase, and clause.

¶49 Supposing there is a demand for high quality translation at a limited budget inadequate for any professional translator, a possible solution is to resort to some online MT service, free or at an acceptably low cost, and then post-edit its translation outputs.⁸³ This post-editing process, which has been involved in human-aided machine translation (HAMT) for a considerable time, is a constructive way to benefit from both the efficiency of machines and the precision of human work. It certainly helps to reduce the total cost, including time and money, for achieving a desired translation quality. The HAMT approach is currently a common practice within the European Union.⁸⁴ It is thus reasonable to expect more users to benefit from this approach with the aid of so many freely available online MT systems. Without a doubt, the starting point for this practice is that the users are provided with reliable information about which MT systems could be their best choices of pursuing their purposes of translation. We have endeavored in this study to provide legal translators, law librarians, and law library users with the most reliable information of this kind so far about a number of popular online MT systems suitable for legal translation.

¶50 Our evaluation provides evidence that Babel Fish can achieve BLEU scores of between 0.2062 and 0.2807—twice as much as stated by Somers⁸⁵—for translating large-scale legal texts from these two languages into English, showing how severely Babel Fish’s translation performance could be underestimated if using a tiny test data set of twenty sentences. In addition to using large-scale data sets and eliminating the problems inherent in human judgment, our evaluation involves two evaluation metrics, six representative online MT systems, fourteen major languages, and about a hundred translation tests. Both the state-of-the-art methodology and the scale of our evaluation ensure that our evaluation results are

82. White, *supra* note 39, at 230–31.

83. See Hutchins, *supra* note 26, at 28.

84. See *id.* at 11.

85. Somers, *supra* note 7, at 616, n. 29.

more reliable and more indicative of their real performance than a subjective manual evaluation with only twenty sentences in two language pairs. Moreover, our evaluation is not about whether a particular MT system can translate well enough. Instead, it is aimed at finding out which (of the best few ones) can translate best for which language pair. This kind of information is certainly most needed by potential users.

¶51 As a last word, we would say that whether an existing MT system is useful or not depends not only on how well it can translate but also largely on how it is utilized. When there is real demand for translation and the suitability, strengths, and weaknesses of available MT systems are well understood, why not incorporate online MT services into one's working environment, especially for legal translation? More practically, they are free of charge.