Modern MT Systems and the Myth of Human Translation: Real World Status Quo.

Richard Jelinek
The PetaMem Group
Germany/Czech Republic
rj@petamem.com

Translating and the Computer 26 Conference, 2004-11-18

Abstract

This paper objects to the current consensus that machine translation (MT) systems are generally inferior to human translation (HT) in terms of translation quality. In our opinion, this belief is erroneous for many reasons, the both most important being a lack of formalism in comparison methods and a certain supineness to recover from past experience. As a side effect, this paper will provide evidence for a much more favorable judgment of the performance of contemporary MT systems. We will present and discuss known methods of automatic MT evaluation, give real world examples of both machine and human translation and finally suggest an universal formal evaluation method to handle both human, as well as MT output in a comparable fashion.

Keywords:
MT, HT, Machine Translation, Human Translation, Evaluation Methods

1 Introduction

"Shakespeare is good in English, but you need to read him in the original Klingon ..."

The intention of this paper is neither to blandish the quality of machine translation (MT), nor to badmouth the quality of human translation (HT). This is important to keep in mind, as it may occasionally look like that.

This paper serves two main purposes: One is to adjust the perception of MT to HT comparisons, the other is to propose a possible framework for automatized evaluation of translation output - generated either by MT or HT.

Today, if you ask some person at least remotely familiar with the topic of MT, their resume about it will most probably be something like "MT is fast, but in terms of quality

nowhere near the quality of HT". Depending on their past experience and current interests, people will judge the quality of MT output from "Totally useless" over "Sufficient to get a grasp of the translated text" to "Good enough for intended purpose X".

As soon as it comes to judgment of the quality of HT only, things become more interesting. The reader might commemorate here ads from various translations agencies or freelance translators every one claiming to be better than the others - in some rather unspecified way. Also when it comes to judgment of an existing translation by some translator who hasn't done this translation, one might sometimes hear not very complimentary comments. So while there is this claim of superior quality of HT, there are also some inconsistencies with this claim.

Once you try to put objective measures on both MT and HT, the question about proper evaluation methods pops up. This paper also gives an overview of a framework for an universal evaluation method, that attempts to provide an unified and open evaluation interface for both MT as well as HT. To the best of the authors knowledge, there is no existing automatic method or process to operationally evaluate HT. In every known case, evaluation is done by "judgment" of the translation by another human with a similar (or hopefully better) comprehension for the target language in question.

While there are several MT evaluation methods, all rely on one or more existing reference HTs that are considered to be some "axiomatic truth". While this may seem as the best solution currently available, it might very well prove to be the primary reason for problems that traditional MT evaluation methods still encounter.

1.1 The Need for Unified MT/HT Evaluation

The need for and purpose of automatic MT evaluation is discussed in several papers such as [6] or [2]. We could summarize these as:

- The need of MT systems creators to monitor the progress of their products.
- The need of MT system users to compare various products.
- To lower the required ressources (cost and time) compared to an evaluation of MT systems performed by humans.

While all of these also represent desirable goals and formidable achievments have been made to reach these, there are some problems with the current - restricted - approach:

- The research focus today is geared mainly to MT output. Evaluation of HT seems not to be an issue.
- Most MT evaluation methods require one or more reference HT.
- The majority of the current methods are basically document similarity methods or enhanced versions hereof.

Consider the advantages of an evaluation method that could evaluate both MT as well as HT output without being dependent on a reference translation:

- Evaluation without having a reference translation would be possible, moreover eliminating the axiomatic position of reference translation in todays evaluation systems. Finally, the removal of the need for a reference translation would further lower the cost of the evaluation process itself.

- Fast, reliable and low-cost comparisons of HT/HT translations would enable companies to quickly decide which translator to choose.
- A quantitative evaluation of MT/HT comparisons would give the opportunity to both MT vendors as well as translators to mark their position in the translation business.

The PetaMem Corporation is building a NLP Portal at `http://nlp.petamem.com`. One functionality of the portal is to give visitors the possibility to let translate some text either via MT, human-edited MT or HT to some selected target language. Unlike other sites providing a similar functionality, the goal here was to remove the need for an editor who had to proofread the translations and still ensure a high quality level.

As the translations must be rapid and low-cost there is simply no space for a second reference translation, or intervention of several humans on a single translation, that might cost some cents. On the other hand, no user would spend these cents again, if the delivered quality was not sufficient.

An automatic control instance was needed and it simply had to work. The idea for an unified evaluation method was born.

1.2 The HT Parameters

When aiming for automatic i.e. operationalizable evaluation of HT, one must first have a sound definition of HT and for this, one of translation. According to [4] it is

Vorgang und Ergebnis der Übertragung eines Textes aus einer Ausgangssprache in eine Zielsprache.
(Process and result of the transition of a text from a source language to a target language.)

Let there be consensus, that HT is the aforementioned transitional process done by one or more human beings. Is there really such a thing as HT and if so, can it always claim a superiority over MT?

To answer these questions we must reflect on our consensus of HT. Some examples show evidence, that this definition is either incomplete or too generic:

1. Translations of the same text differ in style, quality and extent between various translators.
2. Translations differ across time. What might have been considered a highly qualitative translation some decades ago may be not appropriate today.
3. When is HT "pure"?

The first point reflects the fact that the child (or adult - for that matter) doing its first translation to a second language may perform very bad. Of course not so bad as a human not knowing this second language at all. A native speaker (target language) will probably generate better results than a non-native speaker and a professional translator and native speaker will probably do better than someone, who is a native speaker but otherwise unexperienced with the business and needs of translation. Finally, given two professional translators that are both native speakers for the target language, the text category may be more familiar to one of them - giving him advantage.

Surprisingly, the second point - adequacy of a given translation to contemporary language - is closely related to the parameters "native language" and "text category".

And finally it is needed to point out, that much of what is called HT is in fact machine-aided HT. That aid might not always be obvious like the fully-fledged translation memory or the output of MT systems for orientation. Also the use of dictionaries[1] or even the simple word processing or the internet can be considered as "machine aid". So if we accept that HT does include external work/help, one parameter to judge the potential quality of the HT is the translators ability to spot, choose and adapt this help to augment his work.

2 Present Methods of Evaluation

"Best fish."

2.1 MT Quality Evaluation

Present methods of MT evaluation are statistical in nature, the most relevant being BLEU [6] and RED [1]. We will mention the widespread BLEU here for reference as it is sufficiently representative for our point. The rationale behind BLEU is, that better MT will be more similar to a reference HT than bad MT. BLEU does under certain circumstances a fairly good job evaluating MT, but shows problematic behavior in some cases.

Many of the problems are connected to the fact, that BLEU is heavily depending on N-Gram based similarity metrics which actually makes it a document similarity measuring method. Unfortunately the N-Gram mechanism is taken for granted as working for this application area without being truly understood or explained [5]. Of course this can only lead to the conclusion, that MT evaluation methods cannot be seen as understood or explained also.

Besides this rather technical or definitional problem, there is also a problem with the required reference translations - which could be considered more fundamental. One or more reference translations (performed by human translators) are matched using N-Gram statistics against the output of MT systems. MT-evaluation literature if full of discussion about the technical problems that occur with this approach such as recall, but no one seems to question the validity of the quasi-axiomatic position of the (human) reference translations.

2.2 HT Quality Evaluation

Evaluation of HT quality today is done either via the aforementioned methods (e.g. BLEU) comparing some translation against a reference translation, or manually by some human judging. It is, however, output-centered. What counts is the result of the translation, not the skills of a translator that led to the translation. While this may be a pragmatical approach, to quickly evaluate a translator, it has some serious drawbacks. Good translators might get rejected just because they were to translate a text category that is not within their primary domain. And of course not always a reference translation is accessible.

Finally the question arises what criteria back up a translation so it becomes the "axiomatic reference" as which it is treated. Unfortunately this question remains often unanswered.

2.3 Comparison MT vs. HT

From what we've heard so far, we can state that:

[1] not necessarily in electronic form, as few of todays printed dictionaries are hand-made without machine intervention...

- Whenever someone mentions HT, he really means more likely: "Machine aided human translation of a skilled professional translator who is native speaker of the target language and has experience with translation of texts of the same category/with a similar topic.
- Whenever someone mentions MT, he really means: "An automatic system working with no or little human intervention to translate general purpose text."

Sounds a little bit unfair - doesn't it? So let's turn things the other way round and - when comparing MT to HT - let's compare today's state of the art MT systems (evtl. specialized on some content) with the average human translation. The next section does exactly that. It takes real world examples of HT and compares them with the output of various MT systems.

3 Real World Examples of Translation Quality

"It is a thesaurus in which entry word,
classification number, sub-classification numbers."

3.1 Man or Machine?

Let's look at one of many real world examples the author was able to gather. Guess what's MT and what HT:[2]

```
Original: Einzigartiger Freizeitpark für Groß und Klein

T1: Singular recreational park for large and small
T2: Unique leisure time park for largely and small
T3: Ein Fantastische DinoPark ferrcoitung
T4: Unique Freizeitpark at big and little
T5: Unique amusement park for great and Klein
T6: Unique leisure park for big and little
```

T1 is the result of the Babelfish/SYSTRAN MT, T2 is by SDL FreeTranslation, T3 is a human translation(!), T4 is NeuroTran, T5 is the Linguatex eTranslation Server and T6 is the PetaMem LangSuite MT.

3.2 HT - annotated

Example 1: Zoological Garden Plzen (Pilsen) City

CS: Jedinečný zábavní park pro malé i velké[3]
EN: Ein Fantastische DinoPark ferrcoitung (??)
DE: Einzigartig Freizeitpark für Kindern und Erwachsene[4]
DE2: Auf der Fläche von 3ha es zeigt die Szenen von Erdmittelaltertiere so, wie

die wahrscheinlich unser Planet vor 200-65 Millionen Jahren bewohnt haben. [5]

[2] source language is German, target language English - just in case
[3] An unique amusement park for small and big (ones)
[4] Two Errors as-is, but really should be: "Einzigartiger Freizeitpark für Groß und Klein"
[5] Auf einer Fläche von 3 ha werden Szenen aus dem Leben von Dinosauriern so gezeigt, wie diese wahrscheinlich unseren Planeten vor 200 - 65 mio. Jahren bewohnt haben.

DE3: Zugänglich einzeln oder als der Teil den Zoo Pilsen. [6]

Example 2: Parking notice board
The following is taken from a notice-board written in english, german and czech that is on the parking place of the czech Sybase subsidiary:

Attention Drivers
Please do not leave valuables in your cars.
This is a high risk area.

Beachtung für den Fahrer
In eigenem Interesse lassen Sie nicht bitte
die Wertsachen im Automobil.

Upozornění pro řidiče
Ve vlastním zájmu nenechávejte
cenné věci ve vozidle.

Considering the quality of the texts we could assume, that the source text of these is either english or czech. However, as both differ significantly, and the german text seems to be a weak translation of the czech one and given the fact that the sign was photographed in the Czech Rep., we consider the czech text being original. An english verbatim translation of it reads:

"**Notice to drivers** In your own interest don't leave valuables in the vehicle."[7]

The german verbatim translation could be:

"**Mitteilung an die Autofahrer** Lassen Sie in Ihrem eigenen Interesse keine Wertsachen im Fahrzeug."

Where the form of address of the drivers would probably be: "Achtung Autofahrer"

Example 3: German "Original"
The German original from the first example of the next section was quoted, because we suspect it being not an original, but a German translation of an English original.[8]

3.3 MT from various Systems - annotated

Example 1: Commercial Text DE→EN past and today
The first example is taken from [3]. The text called "German Original" and the 1999 Babelfish/SYSTRAN translation are cited from this paper, the 2004 Babelfish/SYSTRAN translation was initiated 2004-09-11 by the author. A HT of the german text is mentioned in [3].

German "Original"

[6] Einzeln oder als Teil des Zoo Pilsen zugänglich.

[7] The czech text gives no hint, whether "řidiče" is the singular or plural "driver(s)" - we assume plural in this case for the translations.

[8] this was confirmed by Mr. S. Bass

Advanced Language Translation bemüht sich, nur mit akkreditierten Übersetzern zu arbeiten. In Situationen, in denen das nicht möglich ist, werden sie von Advanced Language Translation direkt vor Ort bewertet und getestet. Hierbei handelt es sich um eine Qualitätssicherungsmaßnahme, die Ihrem Unternehmen garantiert, daß sich Advanced Language Translation um hervorragende Leistungen bemüht, wodurch Advanced Language Translation weiterhin den Ruf genießt, Übersetzungen höchster Qualität zu liefern.

Babelfish/SYSTRAN 1999

Advanced LANGUAGE translation strives to operate only with accredited **compilers/translators**. In situations, in which that is not possible, they are directly locally evaluated and tested by Advanced LANGUAGE translation. Here it concerns a quality assurance measure, which guarantees your enterprise that Advanced LANGUAGE translation strives for outstanding **performances**, whereby Advanced LANGUAGE translation enjoys further the call supplying translations of highest quality.

Babelfish/SYSTRAN 2004-09-11

Advanced LANGUAGE translation strives to work only with accredited **translators**. In situations, in which that is not possible, they are directly locally evaluated and tested by Advanced LANGUAGE translation. Here it concerns a quality assurance measure, which guarantees your enterprise that Advanced LANGUAGE translation strives for outstanding **achievements**, whereby Advanced LANGUAGE translation enjoys further the call supplying translations of highest quality.

We see only two changes between the 1999 and 2004 Babelfish translation (marked in bold), that seem to represent minor manual corrections in the underlying vocabulary of Babelfish. Considering the more serious problems with this translation such as the unmotivated all-caps of “Language”, the evident non-recognition of named entities and missing phrasal optimization, the evolution of this MT seems negligible. There seem to be some problems with the dictionary, as the word “Language” is considered a German word whose english translation is “LANGUAGE”.

SDL FreeTranslation

Advanced Language Translation strives accredited to work, only with translators. In situations, in which that is not possible, they are valued and tested of Advanced Language Translation directly on the spot. Herewith it concerns a quality security measure, that guarantees your business that Advanced Language Translation strives for outstanding achievements, by what means Advanced Language Translation enjoys further the call to deliver translations of highest quality.

InterTran

Advanced Language Translation bestirs himself , only at accredited translator to jobs. in situations , in this the not possible am , become she of Advanced Language

Translation direct before place appraises and tested herewith act it himself about a Qualitätssicherungsmaßnahme , the her undertaking avouches , that himself Advanced Language Translation about excellent performances bestirs , through which Advanced Language Translation furthermore the call enjoys , translations extreme quality to deliver.

PetaMem MT

Advanced Language Translation tries to work only with accredited translators. In situations, in which that is not possible, they will be evaluated and tested on location from Advanced Language Translation. Here it acts as quality assurance action, that guarantees your enterprise, that Advanced Language Translation tries for outstanding performances, through which Advanced Language Translation furthermore enjoys the call to deliver translations of highest quality.

Please judge the quality and make your own ranking of the translations. Compare then your results with the quantitative results in section 5.3.

Example 2: MT of example 2 texts of previous section

InterTran - CS→DE:

Warnte seinetwegen verspotten. Unter Land unterziehen Verhaftete Nörgler Wert veci unter vozidle. (*Warned for his sake ridicule. Below land undergo Arrested Nagger Value veci below vozidle.*)

InterTran - CS→EN - with edit:

Warned to(sake) ridice. Under country veci take prisoner nagger worthiness veci in(under) vozidle.

PetaMem MT - CS→DE 1st try:

Aufmerksammachung für Fahrer. In eigenem Interesse lassen Sie nicht wertvolle Sachen im Fahrzeug. (*Advise-making for driver/drivers. In own interest do not let valuable things in the vehicle.*)

PetaMem MT - CS→DE 2nd try: [9]

Hinweis für Fahrer. In eigenem Interesse lassen Sie nicht Wertsachen im Fahrzeug.(*Advice for driver/drivers. In own interest do not let valuables in [the] vehicle.*)

4 Evaluation Revisited

eval(*"Best fish."*)

In this section we want to summarize known problems with both MT and HT evaluation, give reasons for these problems and finally to make suggestions how to cure or at least circumvent them. This all should make us ready for a proposal of a generic evaluation framework that can cope with both MT and HT.

[9] Adding "Upozornění" to lexicon instead of letting the system derive it from "upozornit" and added peephole optimization in DE "wertvolle Sachen/Wertsachen"

4.1 Known Problems with MT Evaluation

1. MT evaluation methods are not really quality evaluation methods, but document similarity evaluation methods (ngram/edit distance). As for ngram-based methods, the underlying mechanism is not even fully understood.

2. There is need for one or more reference translations (most likely HT). With only one reference translation a bias in translation evaluation is imposed, with several reference translations problems with recall arise.

3. Reference translations are expensive and may not even be available in certain situations.

4. Reference translations are considered a kind of "axiomatic truth" and their evaluation virtually does not happen. They are "judged".

4.2 Known Problems with HT Evaluation

It is our belief, that there are exactly the same problems with HT evaluation as those with MT evaluation. Paradoxically, in literature there are no known problems with HT evaluation, probably because it is considered "working" as of now (by "judging"). A formal - preferably automatic - evaluation of HT seems to be a non-issue, even absurd.

Nevertheless, evaluating HT the same way as MT could be done - and has been done using regular MT evaluation methods (see [5]). The fact, that during these evaluations some MT results scored better than their HT counterparts is certainly not because of the "real quality" of these, but because of the insufficient evaluation method and its modeling. Because conventional evaluation methods are document similarity evaluation methods, they punish "free translations" that have quite different wording, but may have the correct semantics and even a better style.

4.3 Information Sources & Tools

Letting all the known problems with MT and HT evaluation aside for a while - what would be needed if we at least tried an attempt at formal translation evaluation (be it MT or HT)? In every case the information available is the source and target text.

What information can we get from the texts? This depends on our available tools for analysis.

- Given some low-level text processing algorithms, we can obtain basic metrics. Length, word/sentence/paragraph count, average word/sentence/paragraph length.

- Given a statistical processing package, we can compute char/word occurence, ngram distribution, etc.

- Given suitable monolingual corpora for both source and target language, we would have statistical "reference values" for these. Together with the statistical package this already allows us to perform basic language identification and text categorization operations on the source and target text.

- Given voluminous SL↔TL dictionaries and monolingual thesauri, we can perform a simple adequacy check and translation distance. With this we are also able to obtain sentence-alignment between source and target text.

Representative voluminous corpora, dictionaries and thesauri seem to perform better as some kind of "axiomatic truth" reference. And in fact they do serve us well. Higher level functionality like morphosyntactical parsing and semantic inference. can provide additional evaluation well beyond the possibilities of conventional evaluation methods.

5 MT ↔ HT Comparison Proposal

"There's no problem in comparing apples to pears."

With the data and toolset mentioned in the previous section, we are able to build a framework for evaluation that is independent on some reference translation and can be used to quantitatively evaluate both MT as well as HT.

5.1 Process Description

The components of the evaluation process and their function are described as follows:

Language identification:
While a client does know the source language, he may not even be able to read the translated text in the target language. A language identification step does provide verification of this fact. Language identification is a very useful application of existing document similarity methods - e.g. utilizing ngrams. With modern methods, the language and encoding of the target text can be verified.

Length comparison:
A very simple but also very rough indication of translation adequacy is length comparison. Parallel corpora do provide information about length of equivalent texts. Exceeding some deviation threshold could be a hint for inadequacy of the translation. This test can run - depending on the length of the text - at word or sentence level.

Text categorization:
An integrated utility providing an identical set of categories across several supported languages can help to match the categories of the source and target texts. A difference of the text categories is also a hint for inadequacy of the translation.

Spellchecking:
Spellchecking of the target text can reveal a badly written text. While spellchecking is mandatory on the translators side, it might have been omitted or the spellchecker available to the translator might not be of sufficient quality. Moreover, this step can very easily integrate checking of requested nomenclature.

Morphosyntactic Analysis/Parsing:
Until now, all mentioned methods could only prevent from serious errors on the translators side, such as sending some wrong text (which was meant for another client). But in case the language, length, category and spellchecking are ok, we still do not know anything about the quality of the submitted text. A morphosyntactic parser for the target language can reveal errors beyond the scope of the aforementioned statistical methods. While each of these itself is not sufficient to guarantee to pass only qualitative translations, evaluation of real-world translations show very good results for rejection of unsatisfactory work.

Semantic Analysis:

Although statistical methods and morphosyntactical analysis itself provide sufficient results, texts can be construed to pass such checking. To prevent these manipulations[10], it is inevitable to resort to a linguistic knowledge base.

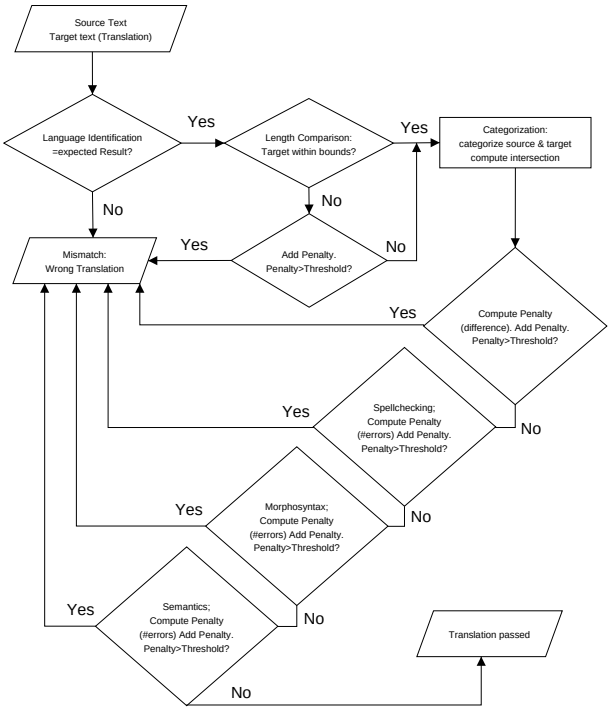

Figure 1: Top-view workflow of MT/HT evaluation process

Figure 1 sketches the basic process workflow, where the available NLP tools are applied to the target text (and source in case of categorization). Except from language identification, which is a k.o. criterium, any deviance from expected or ideal results is added as penalty that cannot exceed some given threshold for the translation to succeed in this evaluation.

We should finally note, that the current implementation of the semantic inference engine (analysis) is very rudimentary in a way that is checking only basic terms for truth or adequacy. Terms it does not match/understand are considered ok.

5.2 Validation on Real World Examples

E1-OE Original english text (non-native speaker):

> Furthermore the look&feel of the portal can be designed to match almost any taste or need thus enabling the portal to seamlessly being integrated into existing solutions with a specific look. Also, the portal can take a specific look when referred from a special site giving partners the possibility to extend their websites by providing NLP functionality through a redirect.

E1-HG Hastily translation to german by the same author (native speaker):

[10] actually it is not a guaranteed prevention, but only further hardening of manipulation

Weiterhin kann das Aussehen des Portals derart frei gestaltet werden, so daß es nahezu jedem Geschmack und Bedürfnis angepasst werden kann und somit dem Portal eine nahtlose Integration in bestehende Lösungen mit einem bestimmten Aussehen ermöglicht. Darüberhinaus kann das Portal ein spezifisches Aussehen und Verhalten annehmen, wenn es von einer bestimmten Seite referenziert wurde. Damit besteht für partner die Möglichkeit ihre Webseiten bei gleichbleibendem Aussehen mit NLP Funktionalitäten durch einfache Referenz zu erweitern.

E1-ME Babelfish translation to english of the german text:

Further the appearance of the portal can be arranged in such a manner freely, so there it almost any taste and need be adapted can and thus the portal a smooth integration into existing solutions with a certain appearance made possible. In addition the portal can take a specific looking and holding back, if it were referenziert from a certain page. Thus the possibility exists to extend their web pages with continuous appearance with NLP functionalities by simple reference for partners.

E1-MG Babelfish translation to german of the english text:

Ausserdem kann das look&feel des Portals entworfen werden, um fast jeden möglichen Geschmack zusammenzubringen oder dem Portal zu integriert werden in vorhandene Lösungen folglich seamlessly ermöglichen zu müssen mit einem spezifischen Blick. Auch das Portal kann einen spezifischen Blick nehmen, wenn es von einem speziellen Aufstellungsort verwiesen wird, der Partnern die Möglichkeit gibt, um ihre Web site zu verlängern, indem es NLP Funktionalität durch umadressieren zur Verfügung stellt.

We can use these texts to cross-validate our evaluation method using E1-OE and E1-HG as source texts and E1-ME and E1-MG as target texts. Furthermore we can use E1-OE and E1-HG as source/target pairs vice versa.

5.3 Quantitative and Qualitative Analysis

Let's have a look at the translations from the previous section and presume these are translations of a source text we consider being allright and that we know nothing of the target language. Table 1 shows the quantitative results of the two german translations, table 2 shows the quantitative results of the same process applied to the english texts when these are considered translations of E1-HG.

Table 1: Evaluation of E1-OE translations

src: E1-OE	**E1-HG**	**E1-MG**
LangIdent	ok(de)	ok(de)
LenComp	!+13(1.2167)	+8(1.1333)
TextCat	3/4	2/4
Spellchk	fail(1)	fail(5)
Parsing	ok(0)	fail(2)
SemStat	ok(0)	fail(3)

In Table 1 we see E1-HG performing better than E1-MG in nearly all tests. It is, however, considered to have excess length, which probably is true as it contains additional statements and information not occuring in the original text.

For Text Categorization, we us a proprietary implemetation with a taxonomy scheme similar to that of USAS[11]. It consists of a well balanced hierarchical discourse structure and is trained on both english and german texts (based on parallel corpora). Currently our matching scheme is limited to a simple lowest-level category match

Table 2: Evaluation of E1-HG "translations"

src: E1-HG	**E1-OE**	**E1-ME**
LangIdent	ok(en)	ok(en)
LenComp	!−13(0.8219)	−8(0.9012)
TextCat	2/4	1/4
Spellchk	fail(1)	fail(1)
Parsing	ok(0)	fail(2)
SemStat	ok(0)	fail(2)

Table 2 shows evidence, that the process currently does not yield identical results when switching source and target texts. The problem is that the trained categorizer is naturally not working exactly on parallel texts in different languages. While this would be desirable, it is hard to achieve. We can also see E1-ME performing better than E1-MG. Assuming a similar degree of difficulty of E1-OE and E1-HG, this could indicate that Systran/Babelfish performs better when the target language is english.

Let us look at a less hastily and more verbatim human translation (E2-HG) of E1-OE and its evaluation in table 3 - we have kept the evaluation results from above for easier comparison:

> Weiterhin kann das Aussehen des Portals so entworfen werden, daß es nahezu jeden Geschmack trifft und ermöglicht somit dem Portal eine nahtlose Integration mit spezifischem Aussehen in bestehende Lösungen. Auch kann das Portal ein spezifisches Aussehen annehmen wenn es von einer bestimmten Webseite referenziert wurde und gibt so Partnern die Möglichkeit durch einen Verweis ihre Websiten mit NLP Funktionalität zu erweitern.

Table 3: Evaluation of E1-OE translations

src: E1-OE	**E1-HG**	**E1-MG**	**E2-HG**
LangIdent	ok(de)	ok(de)	ok(de)
LenComp	!+13(1.2167)	+8(1.1333)	+1(1.0167)
TextCat	3/4	2/4	3/4
Spellchk	fail(1)	fail(5)	ok(0)
Parsing	ok(0)	fail(2)	ok(0)
SemStat	ok(0)	fail(3)	ok(0)

[11]UCREL Semantic Analysis System - see `http://www.comp.lancs.ac.uk/ucrel/usas/`

Although evaluation of E2-HG shows better results for this translation, it becomes evident, that this automatic evaluation method cannot sufficiently account for style. E2-HG seems a little bit dry even clumsy of a german text. Then again, E1-OE was not created by a native speaker, so we could interpret this result also in a way, that the E1-OE is probably dry and even clumsy of an english text.

Finally, let us have a look at the evaluation of the various MT output from section 3.3 and compare them in table 4. Source text (S1) is the German "Original", T1 is Systran 1999, T2 is Systran 2004, T3 is SDL FreeTranslation, T4 is InterTran, T5 is PetaMem MT, T6 is the HT from [3], T7 is HT done by a german native speaker who "never was that good in english" and who was asked to perform the translation as fast and as good as possible. The translator has not seen any of the translations prior to translating text to:

> Advanced Language Transaltion bestirs itself, only working with accredited translators. In situations in which that is not possible, they will directly be evaluated at place by Advanced Language Translation and tested. At this it is a matter of a quality assurance action, that is guaranteeing your company, that Advanced Language Translation bestirs itself, whereby Advanced Language Translation furthermore enjoys the reputation, to deliver highest quality Translations.

Text was cut&paste from email, underlined words were looked up in a dictionary by the translator.

Table 4: Comparison of MT & HT translations

S1	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
LangIdent	ok(en)	ok(en)	ok(en)	ok(en)	ok(en)	ok(en)	ok(en)
LenComp	−3(0.9531)	−2(0.9688)	+2(1.0313)	!+8(1.125)	0(1)	−8(0.875)	+2(1.0313)
TextCat	3/3	3/3	2/3	0/3	3/3	3/3	2/3
Spellchk	fail(1)	ok(0)	ok(0)	fail(2)	ok(0)	ok(0)	fail(2)
Parsing	fail(1)	ok(0)	fail(1)	fail(6)	ok(0)	fail(2)	fail(2)
SemStat	ok(0)	ok(0)	fail(1)	fail(3)	ok(0)	ok(0)	fail(1)

The reader may compare the results with the judgement he made at end of section 3.3 and may or may not find similarities at least with the ranking that the values in table imply. While it is clear, that the T6 HT represents the best translation, T7 HT is outperformed by T5 and T2. It should also be noted, that - as for length comparison - T6 is hard on the limit for DE:EN translations. A careful inspection also shows, that the german text contains more information than the english translation (in fact it is vice versa).[12]

6 Concluding Remarks

"Famous last words:"

This paper has shown evidence, that the best MT systems are far better than the worst human translators. While this might not seem as a surprise right now, it is certainly important to keep this differentiated view in mind when making general or comparative statements about the quality of both human and machine translation.

[12]The failures in parsing are from missing interpunction in the english translation and would not be present with correct commas.

The most significant things to consider and to remember should be:

- When referring to HT, people often mean "the best HT available". The real-world HT diaspora is of course not bounded below.
- While also not bounded below, contemporary MT systems tend to be of great qualitative diversity also. Before rejecting the usage of "a MT system" or general statement MT being worse than HT, one should keep in mind, that not few of todays translations that are actually **done** by humans are worse or equal to translations performed by machines.

We have shown a framework for automatic translation evaluation (both MT and HT). This framework consists of a combination of various NLP/NLU functionality, most of which is commodity today. One could argue, that the framework does no real evaluation, but only validation of a given translation. While we agree, that this proposal for automatic evaluation is just a start with many details that need to be worked upon, the system has already proven to be suitable for everyday use in MT/HT control. It does not guarantee to pass only qualitative translations, however it does reliably reject bad translations.

The inability of the system to make statements on translation style currently does not mean much of a drawback. From our experience, most project managers or supervisors in translation agencies also just control whether the correct translation is sent and whether it is matching certain criteria - mostly formatting and spellchecking.

We are currently investigating the usability of large monolingual corpora and a phrasal evaluation to make statements on style. In the meantime, a semi-automatic evaluation system based on feedback/voting on translation quality will be deployed.

References

[1] Y. Akiba, K. Imamura and E. Sumita. 2001. *Using multiple edit distances to automatically rank machine translation output*. In Proc. MT Summit VIII, pp. 15-20.

[2] B. Babych and A. Hartley. 2004. *Extending the BLEU MT Evaluation Method with Frequency Weightings*, ACL04, pp. 621-628.

[3] S. Bass. 1999. *Machine vs. Human Translation*. Advanced Language Translation, Inc., `http://www.advancedlanguage.com/articles/machine_vs_human_translation.pdf`.

[4] H. Bußmann. 2002. *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, Stuttgart.

[5] C. Culy and S. Z. Riehemann. 2003. *The Limits of N-Gram Translation Evaluation Metrics*. SRI International, MT Summit IX.

[6] 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.