Comparing machine translation and human translation: A case study

Lars Ahrenberg

Conference article

Cite this conference article as:


Copyright: The authors

The self-archived postprint version of this conference article is available at Linköping University Institutional Repository (DiVA):
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-141902
Comparing Machine Translation and Human Translation: A Case Study

Lars Ahrenberg
Department of Computer and Information Science
Linköping University
lars.ahrenberg@liu.se

Abstract
As machine translation technology improves comparisons to human performance are often made in quite general and exaggerated terms. Thus, it is important to be able to account for differences accurately. This paper reports a simple, descriptive scheme for comparing translations and applies it to two translations of a British opinion article published in March, 2017. One is a human translation (HT) into Swedish, and the other a machine translation (MT). While the comparison is limited to one text, the results are indicative of current limitations in MT.

1 Introduction
In the CFP for this workshop it is claimed that ‘Human translation and Machine Translation (MT) aim to solve the same problem’. This is doubtful as translation is not one thing but many, spanning a large space of genres, purposes, and contexts.

The aim of MT research and development is often phrased as ‘overcoming language barriers’. To a large extent this aim has been achieved with many systems producing texts of gisting quality for hundreds, perhaps even thousands of language pairs, and (albeit fewer) systems that enable conversations between speakers that do not share a common language. Human translation, however, often has a more ambitious aim, to produce texts that satisfy the linguistic norms of a target culture and are adapted to the assumed knowledge of its readers. To serve this end of the market, MT in combination with human post-editing is increasingly being used (O’Brien et al., 2014). The goals for MT have then also been set higher, to what is often called quality translation, and new ‘interactive’ and/or ’adaptive’ interfaces have been proposed for post-editing (Green, 2015; Vashee, 2017). Thus, when production quality is aimed for, such as in sub-titling or publication of a news item or feature article, human involvement is still a necessity.

Some recent papers claim that MT now is almost ‘human-like’ or that it ‘gets closer to that of average human translators’ (Wu et al., 2016). While such claims may be made in the excitement over substantial observed improvements in a MT experiment, they raise the question (again!) of how HT may differ from MT.

Some scholars have argued that MT will never reach the quality of a professional human translator. The limitations are not just temporary, but inherent in the task. These arguments are perhaps most strongly expressed in (Melby with T. Warner, 1995). More recently, but before the breakthrough of NMT, Giammarresi and Lapalme (2016) still consider them valid. As MT can produce human-like translations in restricted domains and is increasingly being included in CAT-tools, they insist that MT is posing a challenge for Translation Studies.

In this paper I report a small case study, a close comparison of a human translation and a machine translation from a state-of-the-art system of the same source text. This is done with two purposes in mind. The first concerns how the differences should be described, what concepts and tools are useful to make such a comparison meaningful and enlightening. The second is to assess the differences between state-of-the-art MT and HT, with the caveat, of course, that one pair of translations cannot claim to be representative of the very large translation universe.
2 MT and Translation Studies

The two fields of MT and Translation Studies (TS) have developed separately for almost as long as they have existed. In the early days of both disciplines, some researchers attempted to account for translation in more or less formal linguistic terms, potentially forming a foundation for automatization, e.g. (Catford, 1965). The ‘cultural turn’ in TS moved the field away from linguistic detail and further apart from MT. The 1990:ies saw a common interest in empirical data, but while corpora, and parallel corpora in particular, were collected and studied in both fields, they were largely used for different purposes. For example, it seems that the empirical results generated by TS studies on translation universals (Baker, 1993) did not have much effect on MT.

A problem related to this challenge is that MT and TS lack common concepts and terminology. MT prefers to speak in terms of models, whereas TS is more comfortable with concepts such as function and culture. There is a mutual interest in translation quality assessment (TQA), however, and large-scale projects on MT tend to have some participation of TS scholars. For example, one result of the German Verbmbil project is the volume Machine Translation and Translation Theory, (Hauenschild and Heizmann, 1997) that contain several studies on human translation and how it can inform MT. It is also true of more recent projects such as QTLaunchPad where evaluation of translation quality was in focus, and CASMACAT where the design of a CAT tool was informed by translation process research (Koehn et al., 2015).

Error analysis is an area of common interest. (O’Brien, 2012) showed that error typologies and weightings were used in all eleven translation companies taking part in her study. It was also shown that some categories occurred in all or the large majority of the taxonomies. She concludes though that error analysis is insufficient and sometimes outright inappropriate. This is so because it doesn’t take a holistic view of the text and its utility and paying too little attention to aspects such as text type, function or user requirements. A number of alternative evaluation models including usability evaluation, ratings of adequacy and fluency, and readability evaluation are proposed.

In the MT context the merits of error analysis is that it can tell developers where the major problems are, and users what to expect. A taxonomy which has been popular in MT is (Vilar et al., 2006). To avoid the necessity of calling in human evaluators every time an error analysis is to be performed there have also been work on automatic error classification (Popović and Burchardt, 2011). While simply counting errors seems less relevant for comparing machine translation to human translation, showing what type of errors occur can be useful. We must recognize then that the categories could vary with purpose.

Another line of research studies the effects of tools and processes on translations. This field is quite underresearched, though see for instance (Jiménez-Crespo, 2009; Lapshinova-Koltunski, 2013; Besacier and Schwartz, 2015) for some relevant studies.

2.1 Comparing Translations

The most common standard for comparing translations is probably quality, a notion that itself requires definition. If we follow the insights of TS, quality cannot be an absolute notion, but must be related to purpose and context. For instance, Mateo (2014), referring to (Nord, 1997) defines it as “appropriateness of a translated text to fulfill a communicative purpose”. In the field of Translation Quality Assessment (TQA) the final outcome of such a comparison will then be a judgement of the kind 'Very good', 'Satisfactory', or 'Unacceptable' where at least some of the criteria for goodness refer to functional or pragmatic adequacy (Mateo et al., 2017).

In MT evaluation, which is concerned with system comparisons based on their produced translations, the judgements are more often rankings: 'Better-than' or 'Indistinguishable-from'. One focus has then been on developing metrics whose ratings correlate well with human ratings or rankings. This line of research got a boost by Papineni et al. (2002) and has since been an ongoing endeavour in the MT community, in particular in conjunction with the WMT workshops from 2006 onwards. Most measures developed within MT rely on reference translations and give a kind of measure of similarity to the references.

While judgements such as Good or Unacceptable are of course very relevant in a use context, a comparison of MT and HT may better focus on characteristic properties and capabilities instead. The questions that interest me here are questions
such as: What are the characteristics of a machine-translated text as compared to a human translation? What can the human translator do that the MT system cannot (and vice versa)? What actions are needed to make it fit for a purpose?

Many works on translation, especially those that are written for presumptive translators, include a chapter on the options available to a translator, variously called strategies, methods or procedures. A translation procedure is a type of solution to a translation problem. In spite of the term, translation procedures can be used descriptively, to characterize and compare translations, and even to characterize and compare translators, or translation norms. This is the way they will be used here, for comparing human translation and machine translation descriptively. This level of description seems to me to be underused in MT, though see (Fomicheva et al., 2015) for an exception.

With (Newmark, 1988) we may distinguish general or global translation methods, such as semantic vs. communicative translation, that apply to a text as a whole (macro-level) from procedures that apply at the level of words (micro-level), such as shifts or transpositions. In this paper the focus is on the micro-level methods.

3 A Case Study

3.1 The Approach

The analysis covers intrinsic as well as extrinsic or functional properties of the translations. The intrinsic part covers basic statistical facts such as length and type-token ratios, and MT metrics. Its main focus, however, is on translation procedures or the different forms of correspondence that can be found between units. A special consideration is given to differences in word order as these can be established less subjectively than categorizations. The functional part considers purpose and context, but one translation can in principle be evaluated in relation to two or more purposes, i.e., post-editing or gisting.

Catford (1965) introduced the notion of shifts, meaning a procedure that deviates somehow from a plain or literal translation. A large catalogue of translation procedures, or methods, was provided by (Vinay and Darbelnet, 1958) summarized in seven major categories: borrowing, calque, literal translation, transposition, modulation, equivalence, and adaptation. Newmark (1988) provides a larger set. The most detailed taxonomy for translation procedures is probably van Leuven-Zwart (1989) who establishes correspondence on a semantic basis through what she calls archi-transemes.

A problem with these taxonomies is to apply them in practice. For this reason I will only give counts for coarse top level categories and report more fine-grained procedures only in rough estimates. At the top level we have a binary distinction between Shifts and Unshifted, or literal translations. An unshifted translation is one where only procedures which are obligatory or standard for the target language have been used, and content is judged to be the same. Semantic shifts are as far as possible noted separately from structural shifts.

Shifts are identified at two levels: sentences and clausal units. Relations between units are established on the basis of position and content. At the sentence level position in the linear flow of information is usually sufficient to infer a relation. At the clausal level correspondence must take syntactic relations into account in addition to content. As for content we require only that there is some sort of describable semantic or pragmatic relation.

3.2 The Data

The source text is an opinion article published by the Financial Times on March 17, 2017 entitled Why I left my liberal London tribe and written by David Goodhart. It is fairly typical of a British opinion article. Paragraphs are short with only a few sentences, the length of sentences are quite varied, and long sentences tend to be built both syntactically and with insertions, appositions and parataxis. Table 1 (first column) gives an illustration.

The human translation appeared in the June issue of the Swedish magazine Axess. It was translated manually with no other computer aids than interactive dictionaries and the web for fact checking. No specific brief had been issued with the assignment. The translation was published with only minor edits but under a different title.

The machine translation was produced by Google Translate in the middle of June, 2017. According to Google’s web information, translation from English into Swedish was then using NMT1.

Even a non-Swedish speaker can observe that the paragraph shown in Table 1 has the same num-

---

1https://cloud.google.com/translate/docs/languages
I referred to changing my mind as though it were a rational process, in which one audits one’s beliefs every few years and decides to shift ground on Israel/Palestine or the single market. But that’s not how it works. If, like most educated people, you place a high value on moral and intellectual coherence, your views tend to fit together into something like an explicable worldview. And that usually goes along with informal membership of a network like-minded people. Without having to think very hard you know you all broadly favour and oppose the same things.

Table 1: A source paragraph and its two translations.

<table>
<thead>
<tr>
<th>English source</th>
<th>Swedish MT</th>
<th>Swedish HT</th>
</tr>
</thead>
<tbody>
<tr>
<td>I referred to changing my mind as though it were a rational process, in which one audits one’s beliefs every few years and decides to shift ground on Israel/Palestine or the single market. But that’s not how it works.</td>
<td>Jag hänvisade till att ändra mig som om det var en rationell process, där man reviderar sin tro på några år och bestämmer sig för att flytta marken mot Israel / Palestina eller den inre marknaden. Men det är inte hur det fungerar.</td>
<td>Jag refererade till ett byte av uppfattning som om det vore en rationell process, där man granskar sina åsikter med några års mellanrum och beslutar sig för att ändra ståndpunkt ifråga om Israel och palestinierna eller EU:s inre marknad. Men det är inte så det fungerar.</td>
</tr>
<tr>
<td>That’s not how it works. If, like most educated people, you place a high value on moral and intellectual coherence, your views tend to fit together into something like an explicable worldview. And that usually goes along with informal membership of a network like-minded people. Without having to think very hard you know you all broadly favour and oppose the same things.</td>
<td>Om du, som de flesta välutbildade, sätter stort värde på moralisk och intellektuell samstämmighet brukar ens åsikter passa in i något som en förståelig världskänsla. Och det går oftast med informellt medlemskap i ett nätverk av likasinnade människor. Utan att behöva tanka välklamt svårt, kan du dig allihopa och motsätter sig samma saker.</td>
<td>Om man, som de flesta välbuddade, sätter stor värde på moralisk och intellektuell samstämmighet brukar ens åsikter passa in i något som liknar en förståelig världskänsla. Och med den följer vanligtvis ett informellt medlemskap i ett nätverk av likasinnade. Utan att egentligen behöva fundera på saken vet man att alla på det hela taget är för och emot samma saker.</td>
</tr>
</tbody>
</table>

Table 2: Basic statistics for the three texts. The last line states the average absolute value of length differences at the sentence level.

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>MT</th>
<th>HT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paragraphs</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Sentences</td>
<td>86</td>
<td>86</td>
<td>95</td>
</tr>
<tr>
<td>Word tokens</td>
<td>2555</td>
<td>2415</td>
<td>2603</td>
</tr>
<tr>
<td>Characters</td>
<td>13780</td>
<td>13888</td>
<td>15248</td>
</tr>
<tr>
<td>Type-token ratio</td>
<td>2.84</td>
<td>2.56</td>
<td>2.58</td>
</tr>
<tr>
<td>Mean Sent.length</td>
<td>29.7</td>
<td>28.1</td>
<td>27.4</td>
</tr>
<tr>
<td>Avg length diff.</td>
<td>–</td>
<td>2.0</td>
<td>3.2</td>
</tr>
</tbody>
</table>

Table 3: BLEU and TER scores for different sections of the MT, using HT as reference.

<table>
<thead>
<tr>
<th>Section</th>
<th>BLEU</th>
<th>Bleu(1)</th>
<th>Bleu(2)</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unshifted</td>
<td>42.79</td>
<td>69.0</td>
<td>48.7</td>
<td>0.374</td>
</tr>
<tr>
<td>Shifted</td>
<td>16.84</td>
<td>48.2</td>
<td>23.6</td>
<td>0.662</td>
</tr>
<tr>
<td>All</td>
<td>23.27</td>
<td>59.6</td>
<td>30.7</td>
<td>0.621</td>
</tr>
</tbody>
</table>

4 Results
4.1 Basic Statistics
The visual impression of Table 1 indicates that the human translation is longer than the machine translation. This is confirmed when we look at the translations as wholes, the human translation is longer both in terms of number of words and number of characters. In terms of characters the ratio is 1.01 for the MT and 1.11 for the HT. Yet, when the HT is shorter than the source for a given sentence, the difference can be large. The HT also has more sentences, as the human translator has decided to split eight sentences (roughly 9% of all) into two or three shorter ones. Basic statistics for all three texts are shown in Table 2.

MT metrics are especially valuable for comparisons over time. As we only have one machine translation in this study, we limit ourselves to reporting BLEU (Papineni et al., 2002) and TER (Snover et al., 2006). After tokenization and segmentation into clause units of both the MT and the HT translations, using the latter as reference we obtained the results shown in Table 3. Following the analysis of clauses into Shifted and Unshifted (see section 4.4) we also computed these metrics for the two types of segments separately.

Values were computed with the multi-bleu.perl script provided with the Moses system, and tercom.7.25, respectively.
4.2 Monotonicity

By monotonicity we mean information on the order of content in the translation as compared to the order of corresponding content in the source text. Both translations are one-to-one as far as paragraphs are concerned. As noted, the HT is not altogether one-to-one at the sentence level, but at the level of clauses, the similarities are greater: the order is the same with the exception that the HT has added one clause.

To get a measure of monotonicity all corresponding word sequences s (from the source text) and t (from the translation) of the form s=a:b and t=Tr(b):Tr(a) are identified. The number of instances per sentence is noted as well as the number of words that are part of such a source sequence. The degree of monotonicity is expressed as a ratio between the total number of affected source words and all words in the text. The results are shown in Table 4.

<table>
<thead>
<tr>
<th>Word Order Changes</th>
<th>MT Sents Words</th>
<th>HT Sents Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>36 0</td>
<td>15 0</td>
</tr>
<tr>
<td>1</td>
<td>40 125</td>
<td>40 197</td>
</tr>
<tr>
<td>2</td>
<td>9 72</td>
<td>22 203</td>
</tr>
<tr>
<td>3</td>
<td>1 10</td>
<td>5 52</td>
</tr>
<tr>
<td>≥4</td>
<td>- 0</td>
<td>4 110</td>
</tr>
<tr>
<td>Total</td>
<td>86 207</td>
<td>86 562</td>
</tr>
</tbody>
</table>

Table 4: Number of sentence segments affected by a certain number of word order changes.

A total of 61 changes of word order is observed in the MT, related to 207 words of the source text, or 1.5% of all words. Almost all of them are correct, the large majority of them relates to the V2-property of Swedish main clauses. as in (1), but there are also successful changes producing correct word order in subordinate clauses, as in (2), or a Swedish s-genitive from an English of-genitive as in (3). While the system thus has a high precision in its word order changes, there are also cases where it misses out.

(1) Why do we change2 our minds about things? Varför förändrar vi vårt sinne om saker?
(2) and feel that for the first time in my life1 I2 ... och känner att jag för första gången i mitt liv1 ...
(3) the core beliefs1 of modern liberalism2 den moderna liberalismens2 kärnföreställningar1

The human translation displays almost twice as many word order changes, 116, and they affect longer phrases and cover longer distances. Still 4.1% is not a very large share and confirms the impression that the information order in the human translation follows the source text closely. The human translator does more than fixing a correct grammar, however, and also improves the style of the text, for instance as regards the placement of insertions, as in (4), and shifts of prominence, as in (5).

(4) I have changed1 my mind, more slowly2, about...
   MT: Jag har förändrat1 mig, mer långsamt2, om...
   HT: Själv har jag, om än långsammare2, ändrat1, inställning..
(5) Instead I met the intolerance1 of .. for the first time2
   MT: Istället mötte jag den intolerans av den moderna vänster1 för första gången2
   HT: Istället fick jag för första gången2 möta den moderna vänsterns intolerans1

4.3 Purpose-related Analysis

It is obvious, and unsurprising, that the MT does not achieve publication quality. To get a better idea of where the problems are, a profiling was made in terms of the number and types of edits judged to be necessary to give the translation publication quality. Any such analysis is of course subjective, and it has been done by the author, so the exact numbers are not so important. However, the total number and distribution of types are indicative of the character of the translation. A simple taxonomy was used with six types3:

- Major edit; substantial edit of garbled output requiring close reading of the source text
- Order edit; a word or phrase needs reordering
- Word edit; a content word or phrase must be replaced to achieve accuracy
- Form edit; a form word must be replaced or a content word changed morphologically
- Char edit; change, addition or deletion of a single character incl. punctuation marks
- Missing; a source word that should have been translated has not been

The distribution of necessary edits on the different types are shown in Table 4. The most frequent type is ’word edit’ which accounts for more than half of all edits. In this group we find words that ’merely’ affect style as well as word choices that thwarts the message substantially.

3 All categories except ’Major edit’ have counterparts in the taxonomy of Vilar et al. (2006).
Table 5: Frequencies of different types of required editing operations for the MT (one analyst).

<table>
<thead>
<tr>
<th>Type of edit</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major edit</td>
<td>13</td>
</tr>
<tr>
<td>Order edit</td>
<td>24</td>
</tr>
<tr>
<td>Word edit</td>
<td>139</td>
</tr>
<tr>
<td>Form edit</td>
<td>66</td>
</tr>
<tr>
<td>Char edit</td>
<td>13</td>
</tr>
<tr>
<td>Missing</td>
<td>21</td>
</tr>
<tr>
<td>Total</td>
<td>276</td>
</tr>
</tbody>
</table>

A skilled post-editor could probably perform this many edits in two hours or less. However, there is no guarantee that the edits will give the translation the same quality or reading experience as the human translation. Minimal edits using a segment-oriented interface will probably not achieve that. The style and phrasing of the source would shine through to an extent that could offend some readers of the magazine, although most of the contents may be comprehended without problems, cf. Besacier and Schwartz (2015) on literary text. However, for gisting purposes the MT would be quite adequate.

4.4 Translation Procedures: What the MT System didn’t do

While the human translator did not deviate that much in the order of content from the source text, he used a number of other procedures that seem to be beyond the reach of the MT system. Altogether we find more than 50 procedures of this kind in the HT. The most important of these are:

- Sentence splitting. There were eight such splits, including one case of splitting a source sentence into three target sentences. This procedure also comes with the insertion of new material such as a suitable adverb or restoring a subject.

- Shifts of function and/or category. These are numerous; non-finite clauses or NP:s are translated by a finite clause in Swedish, a complete clause is reduced by ellipsis, a relative clause may be rendered by a conjoined clause, adjectival attributes are rendered as relative clauses, an adverb is translated by an adjective or vice versa, and so on.

- Explicitation. Names whose referents cannot be assumed to be known by the readers are explained, e.g. ‘Russell Group universities’ receives an explanation in parentheses. Also, at the grammatical level function words such as och (and), som (relative pronoun), att (complementizer ‘that’) and indefinite articles are inserted more often in the HT than in the MT.

- Modulation = change of point of view. For example, translating ‘here’ and ‘these islands’ in the source by ‘Storbritannien’ (Great Britain).

- Paraphrasing. The semantics is not quite the same but the content is similar enough to preserve the message, e.g. the translation of ‘move more confidently through the world’ is translated as the terser ‘öka ens självsväkerhet’ (increase your confidence).

5 Conclusions and Future Work

Differences between machine translations and human translations can be revealed by fairly simple statistical metrics in combination with an analysis based on so-called shifts or translation procedures. In our case, the MT is in many ways, such as length, information flow, and structure more similar to the source than the HT. More importantly, it exhibits a much more restricted repertoire of procedures, and its output is estimated to require about three edits per sentence. Thus, for publishing purposes it is unacceptable without human involvement. Post-editing of the MT output could no doubt produce a readable text, but may not reach the level of a human translation. In future work I hope to be able include post-edited text in the comparison.

Another topic for future research is predicting translation procedures on a par with current shared tasks predicting post-editing effort and translation adequacy.

Acknowledgments

I am indebted to the human translator of the article, Martin Peterson, for information on his assignment and work process, and to the reviewers for pointing out an important flaw in the submitted version.
References


