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.

The Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT), pages 21–28,
Varna, Bulgaria, Sept 7, 2017
https://doi.org/10.26615/978-954-452-042-7_003

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 Verbmobil 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'Brian, 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 rightout 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 propsed.

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 presumtive 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 architransemes.

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 NMT[1].

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

[1] https://cloud.google.com/translate/docs/languages

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

Table 1: A source paragraph and its two translations.

ber of sentences as the source in both translations (there are 5), that the sentences correspond one-to-one and are quite similar in length. The flow of information is also very similar; shorter units than sentences such as clauses and phrases can be aligned monotonously in both translations with few exceptions.

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

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

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[2]. 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.

Section	BLEU	Bleu(1)	Bleu(2)	TER
Unshifted	42.79	69.0	48.7	0.374
Shifted	16.84	48.2	23.6	0.662
All	23.27	59.6	30.7	0.621

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

[2]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.

Word Order changes	MT		HT	
	Sents	Words	Sents	Words
0	36	0	15	0
1	40	125	40	197
2	9	72	22	203
3	1	10	5	52
≥ 4	-	0	4	110
Total	86	207	86	562

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$_1$ change$_2$ our minds about things?
Varför förändrar$_2$ vi$_1$ vårt sinne om saker?
(2) and feel that for the first time in my life$_1$ I$_2$ ...
och känner att jag$_2$ för första gången i mitt liv$_1$ ...
(3) the core beliefs$_1$ of modern liberalism$_2$
den moderna liberalismens$_2$ kärnföreställningar$_1$

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 changed$_1$ my mind, more slowly$_2$, about..
MT: Jag har förändrat$_1$ mig, mer långsamt$_2$, om..
HT: Själv har jag, om än långsammare$_2$, ändrat$_1$ inställning..
(5) Instead I met the intolerance$_1$ of .. for the first time$_2$
MT: Istället mötte jag den intolerans av den moderna vänster$_1$ för första gången$_2$
HT: Istället fick jag för första gången$_2$ möta den moderna v¨nsterns intolerans$_1$

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 types[3]:

- 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 deletetion 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).

Type of edit	Frequency
Major edit	13
Order edit	24
Word edit	139
Form edit	66
Char edit	13
Missing	21
Total	276

Table 5: Frequencies of different types of required editing operations for the MT (one analyst).

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. 'Rusell 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 självsäkerhet' (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 repertoir 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

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In Honour of John Sinclair*, John Benjamins, Amsterdam and Philadelphia.

Laurent Besacier and Lane Schwartz. 2015. Automated translation of a literary work: A pilot study. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Association for Computational Linguistics, Denver, Colorado, USA, pages 114–122. http://www.aclweb.org/anthology/W15-0713.

John C Catford. 1965. *A Linguistic Theory of Translation*. Oxford University Press, London, UK.

Marina Fomicheva, Núria Bel, and Iria da Cunha. 2015. *Neutralizing the Effect of Translation Shifts on Automatic Machine Translation Ev aluation*, Springer International Publishing, pages 596–607. https://doi.org/10.1007/978-3-319-18111-0_45.

Salvatore Giammarresi and Guy Lapalme. 2016. Computer science and translation: Natural languages and machine translation. In Yves Gambier and Luc van Doorslaer, editors, *Border Crossings: Translation Studies and other disciplines*, John Benjamins, Amsterdam/Philadelphia, chapter 8, pages 205–224.

Spence Green. 2015. Beyond post-editing: Advances in interactive translation environments. *ATA Chronicle* Www.atanet.org/chronicle-on-line/...

Christa Hauenschild and Susanne Heizmann. 1997. *Machine Translation and Translation Theory*. De Gruyter.

Miguel A. Jiménez-Crespo. 2009. Conventions in localisation: a corpus study of original vs. translated web texts. *JoSTrans: The Journal of Specialised Translation* 12:79–102.

Philipp Koehn, Vicent Alabau, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Jesús González-Rubio, Frank Keller, Daniel Ortiz-Martínez, Germán Sanchis-Trilles, and Ulrich Germann. 2015. Casmacat, final public report. http://www.casmacat.eu/uploads/Deliverables/final-public-report.pdf.

Ekaterina Lapshinova-Koltunski. 2013. Vartra: A comparable corpus for analysis of translation variation. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*. Association for Computational Linguistics, Sofia, Bulgaria, pages 77–86. http://www.aclweb.org/anthology/W13-2510.

Roberto Martíinez Mateo. 2014. A deeper look into metrics for translation quality assessment (TQA): A case study. *Miscelánea: A Journal of English and American Studies* 49:73–94.

Roberto Martíinez Mateo, Silvia Montero Martíinez, and Arsenio Jesús Moya Guijarro. 2017. The modular assessment pack a new approach to translation quality assessment at the directorate general for translation. *Perspectives: Studies in Translation Theory and Practice* 25:18–48. Doi 10.1080/0907676X.2016.1167923.

Alan Melby with T. Warner. 1995. *The Possibility of Language*. John Benjamins, London and New York. https://doi.org/10.1075/btl.14.

Peter Newmark. 1988. *A Textbook of Translation*. Prentice Hall, London and New York.

Christiane Nord. 1997. *Translation as a Purposeful Activity*. St Jerome, Manchester, UK.

Sharon O'Brian. 2012. Towards a dynamic quality evaluation model for translation. *The Journal of Specialized Translation* 17:1.

Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia. 2014. *Post-Editing of Machine Translation: Processes and Applications*. Cambridge Scholars Publishing, Newcastle upon Tyne.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Maja Popović and Aljoscha Burchardt. 2011. From human to automatic error classification for machine translation output. In *Proceedings of the 15th International Conference of the European Association for Machine Translation*. Leuven, Belgium, pages 265–272.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.

Kitty M van Leuven-Zwart. 1989. Translation and original: Similarities and dissimilarities, 1. *Target* 1:2:151–181.

Kirti Vashee. 2017. A closer look at sdl's adaptive mt technology. Http://kv-emptypages.blogspot.se/2017/01/a-closer-look-at-sdls-adaptive-mt.html.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of machine translation output. In *LREC06*. Genoa, Italy, pages 697–702.

Jean-Paul Vinay and Jean Darbelnet. 1958. *Stylistique Comparée du Fran cais et de l'Anglais. Méthode de Traduction*. Didier, Paris.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144. Http://arxiv.org/abs/1609.08144.