A Fine-grained Error Analysis of NMT, PBMT and RBMT Output for English-to-Dutch

Laura Van Brussel, Arda Tezcan, Lieve Macken
Ghent University, Department of Translation, Interpreting and Communication
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
laura.vanbrussel@ugent.be, arda.tezcan@ugent.be, lieve.macken@ugent.be

Abstract

This paper presents a fine-grained error comparison of the English-to-Dutch translations of a commercial neural, phrase-based and rule-based machine translation (MT) system. For phrase-based and rule-based machine translation, we make use of the annotated SCATE corpus of MT errors, enriching it with the annotation of neural MT errors and updating the SCATE error taxonomy to fit the neural MT output as well. Neural, in general, outperforms phrase-based and rule-based systems especially for fluency, except for lexical issues. On the accuracy level, the improvements are less obvious. The target sentence does not always contain traces or clues of content being missing (omissions). This has repercussions for quality estimation or gisting operating only on the monolingual level. Mistranslations are part of another well represented error category, comprising a high number of word-sense disambiguation errors and a variety of other mistranslation errors, making it more complex to annotate or post-edit.

Keywords: machine translation, error classification, bilingual corpus

1. Introduction

Since 2016, the landscape of automated translation has substantially changed with the arrival of neural machine translation (NMT). The output quality of this newest system is a hot topic for research at the moment. It has already been compared with the previous state-of-the-art phrase-based machine translation (PBMT) engines and even with rule-based machine translation (RBMT) engines, focusing on the overall performance by applying various automatic metrics, by manual ranking and scoring (Shterionov, Casanellas, Superbo, & O'Dowd, 2017), post-editing or manual error classification (Bentivogli, Bisazza, Cettolo, & Federico, 2016). The scope of the studies range from one to multiple language directions (Toral and Sánchez-Cartagena 2017; Klubička, Toral, and Sánchez-Cartagena 2017; Bojar et al. 2016). Unlike previous work, where engines are developed in research institutes or test suites are built for evaluation, in this paper, we take a different angle by using commercial MT systems and real-life texts from different genres, and thus bring more ecological validity into the field.

In this article, we compare the output of commercial NMT, PBMT and RBMT systems for English to Dutch. Since it provides a detailed overview of the types of errors, we want to discover if the findings for other language pairs apply to English-to-Dutch as well, identify the actual improvements that NMT systems bring to automated translation and get a grip on their potential shortcomings.

2. Related Work

This analysis is carried out in the framework of the SCATE project (Tezcan, Hoste, & Macken, 2017b) and draws on its corpus of PBMT and RBMT errors. We used SCATE's error taxonomy to annotate the same sentences, this time translated by Google's Neural Machine Translation (GNMT)[1].

A substantial part of the research in the field focuses on the language pair English-German. For English to German, Bentivogli et al. (2016) found that NMT output contains less lexical, morphological and word-order errors, which leads to a lower overall post-editing effort. However, according to the authors, the performance of NMT degraded more quickly for longer sentences.

Popović (2017) looked into both the overall performance and the specific language-related issues for German-English, using the output of the best NMT and a PBMT engine which participates in the WMT 2016 shared news translation task. The BLEU score and ChrF-score[2] for NMT were higher than for the PBMT output in both language directions. She manually annotated a subset of 264 sentences for English-to-German and 204 for German-to-English extracted from the total corpus of 3000 sentences. In her study, the number of correct sentences was remarkably higher for the NMT system than for the PBMT system. As for the language-specific issues, NMT outperformed the PBMT system in terms of verb aspects (form, order and omission), articles, English noun collocations and German compounds, as well as phrase structure. This led to improved fluency. Burchardt et al. (2017) use a test suite drawn from grammatical resources, and online lists, consisting of typical translation errors, to compare the output of different NMT, PBMT and RBMT systems. This very controlled, difficulty-isolating method, showed a higher intra-system output variation among NMT systems. They also found that NMT scores best on composition, function words, long-distance dependency, multiword expressions, subordination and verb valence. Ambiguity, tense and mood of verbs, on the other hand, are handled best by RBMT systems. Terminology and named entities, finally, form the mainstay of PBMT systems based on their results. By using a similar challenge-set approach, Isabelle, Cherry, and Foster (2017) focus on short sentences that contain one particular language phenomenon at a time, which reveals the strengths and weaknesses of NMT compared to PBMT for English to French. The controlled input in both studies is both a strength and a trade-off for ecological validity. Language-specific errors hardly ever occur in isolation. The performance of systems can differ if multiple difficulties need to be handled in the same sentence.

[1] MT output generated in June 2017.
[2] Character n-gram F-score (Popovic, 2015)

Toral and Sánchez-Cartagena (2017), Bojar et al. (2016) and Castilho et al. (2017) take more language directions into account to evaluate and compare NMT and PBMT.

The MT systems involved in the news-shared task at WMT 2016 (Bojar et al., 2016) covered the language pairs English to German, Czech, Russian, Finnish, Romanian and Turkish. Thanks to a combination of BLEU scores and human ranking, the output of the best systems could be determined, listing NMT (more specifically the engine submitted by the University of Edinburgh) on top in most language directions or in second place except for English-Finnish. Toral and Sánchez-Cartagena (2017) took the output of the best NMT and PBMT systems from the news translation task of WMT16 as a starting point. They have proven that for all language pairs, there is a higher intersystem variability for NMT output and that the NMT output was more fluent. Furthermore, NMT generates more (correct) word reorderings for almost all language pairs (not the case for EN-DE and EN-FI). A negative correlation between sentence length and performance was confirmed for the majority of language directions. For sentences longer than 40 words, the PBMT systems even outperformed NMT, which they ascribe to the sub-word unit operating level of NMT. Finally, NMT performs better on inflection and reordering for all language directions.

A word of caution has been added by Castilho et al. (2017). For English-German, English-Portuguese, English-Greek and English-Russian professional translators were asked to perform three tasks: post-editing, annotating and ranking the PBMT and NMT output. Although for all language pairs, NMT was ranked as most fluent, NMT produced more correct sentences, contained fewer inflectional and word order errors, and needed less effective post-edits, "the progress is not always evident" they warn. The participants indicated that NMT errors were more difficult to identify, compared to the obvious word-order errors and disfluencies occurring in PBMT output. This is attributable to the higher omission, addition and mistranslation rates for NMT (as opposed to PBMT) in some language directions. They concluded that the throughput and temporal effort only marginally improved thanks to NMT.

3. Error classification

For the annotation and classification of the MT errors, we made use of the SCATE error taxonomy (Tezcan et al., 2017b), which differentiates between fluency (assessing the well-formedness of the target language) and accuracy errors (concerning the transfer of source content).

For more information on the annotation guidelines and process, we would refer to Tezcan, Hoste, and Macken (2017). The advantage of the SCATE annotation method is that both fluency and accuracy errors are annotated separately and that the erroneous MT section is linked to the source section in the case of accuracy errors (except for omissions and additions, which are only labelled in source and target, respectively). The categories in the classification are based on MT-specific errors. A text span can receive multiple labels if different types of errors occur in this span. The fine-grained MT error annotations of SCATE serve as training material to develop Quality Estimation systems for MT (Tezcan, Hoste, & Macken, submitted, 2017a).

For the annotation of NMT, we added two extra categories: i) 'fluency-grammar-extra-repetition' (see section 5.2.2), and ii) 'accuracy-mistranslation-semantically unrelated' (see section 5.1.1). The existing annotations for 'fluency-grammar-extra' in the RBMT and PBMT subsets were revised in case the new category suited the output better. Accordingly, all subsets now bare the same updated annotation labels, allowing for a fair comparison.

4. Research setup

4.1 Data Sets

The SCATE corpus of MT errors was built with sentences extracted from the Dutch Parallel Corpus (Macken, De Clercq, & Paulussen, 2011). From this balanced, commercial and copy-right-cleared corpus, an equal number of 665 sentences was selected from three different text types (non-fiction, external communication and journalistic texts[3]).

4.2 MT systems

For the SCATE corpus of MT errors, created in 2014, Systran[4] was used as RBMT system and Google Translate as PBMT system. Around the beginning of October 2016, Google switched to neural, launching Google's Neural Machine Translation system. The architecture of this model consists of deep Long Short-Term Memory recurrent neural networks (LSTM RNNs) with eight encoder and eight decoder layers that use residual connections and attention connections (Wu et al., 2016).

4.3 Annotations

A total of six annotators (all with a linguistic background) worked on this project. For the PBMT and RBMT subsets, two pairs annotated in parallel in June 2014; in June 2016, one pair was assigned this task for NMT. For the actual annotation process, the brat rapid annotation[5] tool was used. To ensure consistency and a higher inter-annotator agreement, annotation guidelines and a reference translation were provided, together with periodic revision moments for questions and answers.

5. Error analysis

A quick glimpse at the overall error statistics in Table 1 reveals that also for English-Dutch, NMT makes fewer mistakes and generates more sentences that are completely correct.

	RBMT	PBMT	NMT
Accuracy	1309	741	472
Fluency	1831	1531	719
Total	**3140**	**2272**	**1191**

Table 1: Total number of errors

[3] The Dutch Parallel Corpus comprises two additional text types that were not used by SCATE: administrative texts and instructive texts.

[4] Systran Enterprise Edition, version 7.5

[5] http://brat.nlplab.org/

	RBMT	PBMT	NMT
Correct sentences	81	130	217
In %	12%	20%	33%

Table 2 : Correct sentences in MT output

Table 2 illustrates that the NMT output surpasses the other systems. One third of the sentences has been translated correctly by NMT, while this rate is much lower for RBMT and PBMT.

5.1 Accuracy errors

Accuracy concerns the transfer of information and meaning from source to target language. The following main categories can be distinguished: mistranslation, do-not-translate (DNT), untranslated, addition, omission, and mechanical.[6] 'Mistranslations' comprise all errors for which the source content has been translated incorrectly (the subcategories will be mentioned below). The label 'DNT' is used for instances in which one or more source words have been translated unnecessarily, e.g. for proper names. 'Addition' refers to errors in which the target content is not present in the source, while for 'omission' some source content is absent in the target sentence. All mistakes concerning non-meaning (mostly punctuation errors only visible on bilingual level) fall under the category 'mechanical'.

Accuracy errors	RBMT	PBMT	NMT
Mistranslation	972	483	330
DNT	116	14	22
Untranslated	65	69	44
Addition	61	39	2
Omission	43	115	62
Mechanical	52	21	12
Total	**1309**	**741**	**472**

Table 3: Overview of the number of accuracy errors

Table 3 shows that overall, NMT scores better on accuracy than previous systems. However, upon closer inspection, it becomes evident that PBMT handles DNT issues better than NMT. This comes as no surprise, since most of the DNT errors are instances of proper names, a reported strength of PBMT (Burchardt et al., 2017). We also observe that RMBT output contains the fewest omissions. The main category 'mistranslation' is obviously a tough nut to crack for automated translation, as it is the category with the highest number of accuracy errors in all three systems, urging us to dig a little deeper.

5.1.1 Mistranslation errors

'Mistranslation' refers to incorrectly translated source content and is subdivided in the following subcategories: multiword expressions (MWE), part of speech (POS), sense, partial and other. The label 'partial' is used for partial translations of verbs (especially for Dutch, separable verbs). The container 'other' comprises mistranslations of the verb tense and voice, or the number (noun/ verb). To cover the instances for which the target word(s) could never be a plausible translation of the given source word, we introduce the label "semantically unrelated". An example:

> EN: … to build the first ever dynamic billboard to **grace** the streets of Glasgow.
>
> NL: … om het eerste dynamische billboard te bouwen om de straten van Glasgow te **grazen**.

In the sentence above 'grace' is translated by 'grazen', the Dutch equivalent of 'to graze'. This new category reveals a high number of semantically unrelated mistranslations in the NMT output, an error that does not occur in RBMT and only rarely in PBMT output.

	RBMT	PBMT	NMT
MWE	288	139	87
POS	52	44	19
Sense	580	208	117
Partial	4	41	6
Semantically Unrelated	0	9	44
Other	48	42	57
Total	**972**	**483**	**330**

Table 4 : Differentiation of mistranslation errors

Table 4 further illustrates the improvement that NMT has made on almost all mistranslation categories, except for 'other'.

5.1.2 Omissions

Castilho et al. (2017) reported the problem of omissions in NMT output. When scanning the error statistics in Table 5, we can see that also in our data set, NMT makes fewer omission errors than PBMT. However, the ratio of omitted words per omission error is much higher in NMT than in PBMT and RBMT.

	# Omissions	# Annotated words	Average # words per omission
RBMT	43	46	1,07
PBMT	115	125	1,09
NMT	62	93	1,50

Table 5: The number of omission errors compared to the number of words per omission error

Looking back at the corpus, we see that the nature of the omission errors has changed. Often, the NMT output does not provide any clues that source content has been omitted. Wu et al. (2016) already commented: "MT systems sometimes produce output sentences that do not translate all parts of the input sentence – in other words, they fail to completely 'cover' the input, which can result in surprising translations".

[6]The actual SCATE taxonomy also includes the categories 'terminology', 'source' and 'other', but these are left out here as there were no occurrences for any of the three systems.

	RBMT	PBMT	NMT
Total omissions	**43**	**115**	**62**
Content words	6	80	53
Function words	37	35	9
% Content words	0,14 %	69,96 %	85,48 %
Visibility	40	89	19
Invisibility	3	26	43
% Invisibility	7 %	23 %	69 %

Table 6: Subdivision of omission errors based on their type and visibility

For each omission, the number of words, type and visibility was annotated. The type of omission differentiates between content words and function words, the latter having a less severe impact on the accuracy of the translation. NMT drops, on average, more content words than RBMT and PBMT.

Another interesting aspect is the visibility of the omission error when there are traces in the target sentence that source content is missing. The visibility of an omission error can be defined as the expectation of content being missing when only reading the translation (without comparing it to the source). In other words, is the omission visible/ expected at monolingual level? To check for visibility of omission errors, all sentences with omission errors were extracted. The annotator indicated, with yes or no, if it was evident from reading only the target sentences if source content was missing.

A few examples will illustrate the invisibility of the omission error in the MT output when only the Dutch target translation is read. These are all examples from NMT.

EN: In Kinshasa, a Belgian colleague is busy with a similar fistula project with which we **would like to** collaborate more effectively.

NL: In Kinshasa is een Belgische collega bezig met een soortgelijk fistelproject waarmee we effectiever samenwerken.

EN: "There is almost no contact now between Israeli and Palestinian writers," Grossman told **me**.

NL: "Er is nu bijna geen contact tussen Israëlische en Palestijnse schrijvers, " vertelde Grossman.

For RBMT and PBMT, the omissions are being anticipated by fluency errors. The reverse is true for NMT, where the fluency is no indicator that all the source content is being transferred into the target. Even text spans of 4 words are being fluently omitted by NMT. We can conclude that the nature of the omission errors has changed. This forms a major issue not only for annotators and post-editors, but also for everybody using these commercial systems online for free, disposing only of the target text. In research, this issue challenges quality estimation of NMT output and gisting.

5.2 Fluency errors

As mentioned before, fluency deals with the well-formedness of the target language, regardless of the transmission of content and meaning from the source into the target sentence. The following categories are identified: grammar, lexicon, orthography, multiple errors and other. The labels of most error categories are self-explanatory, except for the label 'multiple errors', which is used when an accumulation of fluency errors on a text span makes it hard to identify the error separately. Table 7 gives an overview of the fluency performance of the different systems.

Fluency	RBMT	PBMT	NMT
Grammar	864	932	260
Orthography	290	253	95
Lexicon	533	235	358
Multiple errors	144	110	6
Other	0	1	0
Total	**1831**	**1531**	**719**

Table 7: Overview of the number of fluency errors

As previous research confirms, fluency is handled best by NMT for English-Dutch as well. The improvements are enormous for all categories, except for lexicon, which, therefore, draws our attention.

5.2.1 Lexicon

Lexical errors are split into two subcategories: 'non-existent' and 'lexical choice'. The latter distinguishes 'content words' from 'function words'. Table 8 gives an overview of all the lexical errors in the MT output. From the results in Table 8, it is clear that NMT makes much more lexical choice errors than PBMT, but it is actually only the content words that cause difficulties. For function words, NMT scores best.

Lexical choice	RBMT	SMT	NMT
Total	468	181	304
Content word	290	91	226
Function Word	178	90	78

Table 8: Subdivision of lexical choice errors

In many instances, the category 'Fluency-Lexical Choice' occurs together with an accuracy error for 'Mistranslation-Sense-Content word'. This type of fluency error is the clue that source content has not been rendered correctly.

5.2.2 Grammar

Another well represented fluency error category is grammar. Comparing the three paradigms, we see the progress that has been made. A further subdivision of this category is presented in Table 9.

Grammar	RBMT	PBMT	NMT
Word form	143	245	73
Word order	372	311	42
Extra word(s)	162	99	46
Missing word(s)	162	247	83
Multi word syntax	24	27	11
other	1	3	5
Total	**864**	**932**	**260**

Table 9: Subdivision of grammar errors

It is worthwhile to take a look into 'extra words'. Table 10 shows the newly added 'repetition' subcategory to label words or word groups that are unnecessarily repeated. The rest category 'other' contains all other extra words in the

target sentences, that should not be there. An example of repetition is illustrated below:

> EN: Located above Glasgow Central Station, on the corner of Union Street and Gordon Street, the **55 square metre** LED screen faces directly onto Renfield street -- the second largest retail location in the United Kingdom, and is visible across a range of over 600 metres.
>
> NL: **Het 55 vierkante meter** LED- scherm ligt boven het central station van Glasgow, op de hoek van Union Street en Gordon Street. Het scherm **heeft een oppervlakte van 55 meter** en is direct zichtbaar op Renfield Street, de tweede grootste winkelplaats in het Verenigd Koninkrijk.

The words in bold in the source sentence have been translated in 2 places (also in bold) in the target sentences. A subdivision of all extra word errors is presented in Table 10.

	RBMT	PBMT	NMT
Extra words	162	99	46
Repetition	9	6	15
Other	153	93	31

Table 10: Subdivision of 'grammar extra words' errors

Although NMT has less superfluous words in its output, it has a higher number of repetitions of one or more words than the other systems.

5.3 Long sentences

In literature, long sentences have been reported as a weakness of NMT systems. Bentivogli et al. (2016) found that the performance of NMT degraded faster with increased segment length. Toral and Sánchez-Cartagena (2017) confirmed this negative correlation and even reported that PBMT outperforms NMT for sentences consisting of 40 or more words.

	RBMT	PBMT	NMT
Long sentences			
# Errors	446	335	157
# Target words	1791	1749	1677
# Unique annotated words	821	685	195
% Erroneous words	46%	39%	12%
Short sentences			
# Errors	230	175	104
# Target words	969	958	950
# Unique annotated words	225	228	113
% Erroneous words	23%	24%	12%

Table 11: Performance on long sentences (min. 40 words) compared to short sentences (max. 10 words)

Table 11 shows us that NMT still outstrips PBMT for long sentences. In fact, two of the 38 long sentences in our corpus were translated without errors by NMT. PBMT and RBMT produced no correct long sentences. For the sake of completeness, we include the performance of all engines on all 145 short sentences found in our corpus as well. In addition to the number of errors, Table 11 also presents the number of target words and the number of unique annotated words for each system. In the number of unique annotated words, every erroneous word is only counted once, even though it might be annotated multiple times in the same sentence.

The percentage of wrong words in long and short sentences in our subset for NMT is the same. The expected degradation of NMT performance in long sentences, doesn't hold (anymore). To overcome the accumulation of errors in longer sentences in NMT, different architectures have been examined and tested (Barone, Helcl, Sennrich, Haddow, & Birch, 2017) and all kinds of attentional mechanisms have been investigated (Luong, Pham, & Manning, 2015) and implemented. The GNMT's architecture has also been enhanced by a bi-directional encoder for the bottom layer only, allowing for a maximum possible parallelisation during computation (Wu et al., 2016).

6. Conclusions and outlook

In this paper we compared the NMT output with RBMT and PBMT translations, providing an overview of the strengths and weaknesses of NMT. We explained why we expect that NMT output is more difficult to post-edit, by elaborating on the special and less transparent character of some types of NMT errors. Omissions and mistranslations that are semantically unrelated to the source, will be a future challenge, especially for all activities that only take the translation product into account (e.g. gisting and quality estimation of MT output).

7. Bibliographical References

Barone, A. V. M., Helcl, J., Sennrich, R., Haddow, B., and Birch. (2017). Deep architectures for NMT. In Proceedings of the conference on Machine Translation (WMT), Vol. 1, pp. 99–107, Copenhagen, Denmark, September. Association for Computational Linguistics.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 257–267, Austin, Texas, November.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M. et al. (2016). Findings of the 2016 conference on Machine Translation, In Proceedings of the First Conference on Machine Translation: Shared Task Papers (WMT), Vol. 2, pp. 131–198, Berlin, Germany, August. Association for Computational Linguistics.

Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J.-T., and Williams, P. (2017). A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), pp. 159–170.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., and Way, A. (2017). Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, *108*(1), pp. 109–120.

Isabelle, P., Cherry, C., and Foster, G. (2017). A challenge set approach to evaluating machine translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2486–2496, Copenhagen,

Denmark, September. Association for Computational Linguistics.

Klubička, F., Toral, A., and Sánchez-Cartagena, V. M. (2017). Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), pp. 121–132.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1412–1421). Lisbon, Portugal: Association for Computational Linguistics.

Macken, L., De Clercq, O., and Paulussen, H. (2011). Dutch parallel corpus: a balanced copyright-cleared parallel corpus. *Meta: Journal Des Traducteurs/Meta: Translators' Journal*, *56*(2), pp. 374–390.

Popovic, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation* (WMT), pp. 392–395, Lisboa, Portugal, September. Association for Computational Linguistics.

Popović, M. (2017). Comparing Language Related Issues for NMT and PBMT between German and English. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 209–220.

Shterionov, D., Nagel, P., Casanellas, L., Superbo, R., and O'Dowd, T. (2017). Empirical evaluation of NMT and PBSMT quality for large-scale translation production. In Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT), p. 74, Prague, Czech Republic, May.

Tezcan, A., Hoste, V., and Macken, L. (submitted). Estimating Word-Level Quality of Statistical Machine Translation Output Using Monolingual Information Alone. *Computer, Speech & Language*.

Tezcan, A., Hoste, V., and Macken, L. (2017). A Neural Network Architecture for Detecting Grammatical Errors in Statistical Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), pp. 133-145.

Tezcan, A., Hoste, V., and Macken, L. (2017). SCATE Taxonomy and Corpus of Machine Translation Errors. In G. Corpas Pastor & I. Durán Muñoz (Eds.), *Trends in e-tools and resources for translators and interpreters*, Brill: pp. 219–248.

Toral, A., and Sánchez-Cartagena, V. M. (2017). A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Vol. 1, pp. 1063-1073. Valencia, Spain, April. Association for Computational Linguistics.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv Preprint arXiv:1609.08144*.