How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Abstract

Analysing translation quality in regards to specific linguistic phenomena has historically been difficult and time-consuming. Neural machine translation has the attractive property that it can produce scores for arbitrary translations, and we propose a novel method to assess how well NMT systems model specific linguistic phenomena such as agreement over long distances, the production of novel words, and the faithful translation of polarity. The core idea is that we measure whether a reference translation is more probable under a NMT model than a contrastive translation which introduces a specific type of error. We present LingEval97\(^1\), a large-scale data set of 97 000 contrastive translation pairs based on the WMT English–German translation task, with errors automatically created with simple rules. We report results for a number of systems, and find that recently introduced character-level NMT systems perform better at transliteration than models with byte-pair encoding (BPE) segmentation, but perform more poorly at morphosyntactic agreement, and translating discontiguous units of meaning.

1 Introduction

It has historically been difficult to analyse how well a machine translation system can learn specific linguistic phenomena. Automatic metrics such as BLEU (Papineni et al., 2002) provide no linguistic insight, and automatic error analysis (Zeman et al., 2011; Popovic, 2011) is also relatively coarse-grained. A concrete research question that has been unanswered so far is whether character-level decoders for neural machine translation (Chung et al., 2016; Lee et al., 2016) can generate coherent and grammatical sentences. Chung et al. (2016) argue that the answer is yes, because BLEU on long sentences is similar to a baseline with longer subword units created via byte-pair encoding (BPE) (Sennrich et al., 2016a), but BLEU, being based on precision of short n-grams, is an unsuitable metric to measure the global coherence or grammaticality of a sentence.

To allow for a more nuanced analysis of different machine translation systems, we introduce a novel method to assess neural machine translation that can capture specific error categories in an automatic, reproducible fashion.

Neural machine translation (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015) opens up new opportunities for automatic analysis because it can assign scores to arbitrary sentence pairs, in contrast to phrase-based systems, which are often unable to reach the reference translation. We exploit this property for the automatic evaluation of specific aspects of translation by pairing a human reference translation with a contrastive example that is identical except for a specific error. Models are tested as to whether they assign a higher probability to the reference translation than to the contrastive example.

A similar method of assessment has previously been used for monolingual language models (Sennrich and Haddow, 2015; Linzen et al., 2016), and we apply it to the task of machine translation. We present a large-scale test set of English–German contrastive translation pairs that allows for the automatic, quantitative analysis of a number of linguistically interesting phenomena that have previously been found to be challenging for machine
translation, including agreement over long distances (Koehn and Hoang, 2007; Williams and Koehn, 2011), discontiguous verb-particle constructions (Nießen and Ney, 2000; Loáiciga and Gulordava, 2016), generalization to unseen words (specifically, transliteration of names (Durrani et al., 2014)), and ensuring that polarity is maintained (Wetzel and Bond, 2012; Chen and Zhu, 2014; Fancellu and Webber, 2015).

We report results for neural machine translation systems with different choice of subword unit, identifying strengths and weaknesses of recently-proposed models.

2 Contrastive Translation Pairs

We create a test set of contrastive translation pairs from the EN→DE test sets from the WMT shared translation task.$^2$ Each contrastive translation pair consists of a correct reference translation, and a contrastive example that has been minimally modified to introduce one translation error. We define the accuracy of a model as the number of times it assigns a higher score to the reference translation than to the contrastive one, relative to the total number of predictions. We have chosen a number of phenomena that are known to be challenging for the automatic translation from English to German.

1. noun phrase agreement: German determiners must agree with their head noun in case, number, and gender. We randomly change the gender of a singular definite determiner to introduce an agreement error.

2. subject-verb agreement: subjects and verbs must agree with one another in grammatical number and person. We swap the grammatical number of a verb to introduce an agreement error.

3. separable verb particle: verbs and their separable prefix often form a discontiguous semantic unit. We replace a separable verb particle with one that has never been observed with the verb in the training data.

4. polarity: arguably, polarity errors are under-measured the most by string-based MT metrics, since a single word/morpheme can reverse the meaning of a translation. We reverse polarity by deleting/inserting the negation particle nicht (‘not’), swapping the determiner ein (‘a’) and its negative counterpart kein (‘no’), or deleting/inserting the negation prefix un-.

5. transliteration: subword-level models should be able to copy or transliterate names, even unseen ones. For names that were unseen in the training data, we swap two adjacent characters.

Table 1 shows examples for each error type. Most are motivated by frequent translation errors; for EN→DE, source and target script are the same, so technically, we do not perform transliteration.

All errors are introduced automatically, relying on statistics from the training corpus, a syntactic analysis with ParZu (Sennrich et al., 2013), and a finite-state morphology (Schmid et al., 2004; Sennrich and Kunz, 2014) to identify the relevant constructions and introduce errors. For contrastive pairs with agreement errors, we also annotate the distance between the words. For translation errors where we want to assess generalization to rare words (all except negation particles), we also provide the training set frequency of the word involved in the error (in case of multiple words, we report the lower frequency).

The automatic processing has limitations, and we opt for a high-precision approach – for instance, we only change the gender of determiners where case and number are unambiguous, so that we can produce maximally difficult errors.$^3$

2 http://www.statmt.org/wmt16/

3 If we mistakenly introduce a case error, this makes it easier to spot from local context.
### 3 Evaluation

In the evaluation section, our focus is on establishing baselines on the test set, and investigating the following research questions:

- how well do different subword-level models process unseen words, specifically names?
- sequence-length is increased in character-level models, compared to word-level or BPE-level models. Does this have a negative effect on grammaticality?

#### Table 2: NMT hyperparameters. ‘decoder’ refers to function implemented in Nematus (for BPE-to-BPE) and dl4mt-c2c (for *-to-char).

<table>
<thead>
<tr>
<th>system</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPE-to-BPE</td>
<td>20.1 (21.0)</td>
<td>23.2 (23.6)</td>
<td>26.7 (26.5)</td>
</tr>
<tr>
<td>BPE-to-char</td>
<td>19.4 (20.5)</td>
<td>22.7 (22.6)</td>
<td>26.0 (25.9)</td>
</tr>
<tr>
<td>char-to-char</td>
<td>19.7 (20.7)</td>
<td>22.9 (22.7)</td>
<td>26.2 (26.1)</td>
</tr>
</tbody>
</table>

We trained the BPE-to-BPE system for one week, the BPE-to-char and char-to-char for two weeks, following Lee et al. (2016). Table 3: Case-sensitive BLEU scores (EN-DE) on WMT newest. We report scores with detokenized NIST BLEU (mteval-v13a.pl), and in brackets, tokenized BLEU with multi-bleu.perl.

### 3.1 Data and Methods

We train NMT systems with training data from the WMT 15 shared translation task EN→DE. We train three systems with different text representations on the parallel part of the training set:

- BPE-to-BPE (Sennrich et al., 2016a)
- BPE-to-char (Chung et al., 2016)
- char-to-char (Lee et al., 2016)

We use the implementations released by the respective authors, Nematus\(^6\) for BPE-to-BPE, and dl4mt-c2c\(^7\) for BPE-to-char and char-to-char. dl4mt-c2c also provides preprocessed training data, which we use for comparability.

Both tools are forks of the dl4mt tutorial\(^8\), so the implementation differences are minimal except for those pertaining to the text representation. We report hyperparameters in Table 2. They correspond to those used by Lee et al. (2016) for BPE-to-char and char-to-char; for BPE-to-BPE, we also adopt some hyperparameters from Sennrich et al. (2016b), most importantly, we extract a joint BPE vocabulary of size 89 500 from the parallel corpus. We trained the BPE-to-BPE system for one week, following Sennrich et al. (2016a), and the *-to-char systems for two weeks, following Lee et al. (2016), on a single Titan X GPU. For both translating and scoring, we normalize probabilities by length (the number of symbols on the target side).

We also report results with the top-ranked system at WMT16 (Sennrich et al., 2016a), which is available online.\(^9\) It is also a BPE-to-BPE system, but in contrast to the previous systems, it includes different preprocessing (including trucating), other hyperparameters, additional monolin-

\(^6\)https://github.com/rsennrich/nematus
\(^7\)https://github.com/nyu-dl/dl4mt-c2c
\(^8\)https://github.com/nyu-dl/dl4mt-tutorial
\(^9\)http://data.statmt.org/rsennrich/wmt16ystems/
Table 4: Accuracy (in percent) of models on different categories of contrastive errors. Best single model result in bold (multiple bold results indicate that difference to best system is not statistically significant).

<table>
<thead>
<tr>
<th>system</th>
<th>agreement (category and size →)</th>
<th>noun phrase 21813</th>
<th>subject-verb 35105</th>
<th>verb particle 2450</th>
<th>polarity (negation) insertion 22760</th>
<th>deletion 4043</th>
<th>transliteration 3490</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPE-to-BPE</td>
<td>95.6</td>
<td>93.4</td>
<td>91.1</td>
<td>97.9</td>
<td>91.5</td>
<td>96.1</td>
<td></td>
</tr>
<tr>
<td>BPE-to-char</td>
<td>93.9</td>
<td>91.2</td>
<td>88.0</td>
<td>98.5</td>
<td>88.4</td>
<td>98.6</td>
<td></td>
</tr>
<tr>
<td>char-to-char</td>
<td>93.9</td>
<td>91.5</td>
<td>86.7</td>
<td>98.5</td>
<td>89.3</td>
<td>98.3</td>
<td></td>
</tr>
<tr>
<td>(Sennrich et al., 2016a)</td>
<td>98.7</td>
<td>96.6</td>
<td>96.1</td>
<td>98.7</td>
<td>92.7</td>
<td>96.4</td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>99.4</td>
<td>99.8</td>
<td>99.8</td>
<td>99.9</td>
<td>98.5</td>
<td>99.0</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Subject-verb agreement accuracy as a function of distance between subject and verb.

Table 5: Accuracy (in percent) of models on different categories of contrastive errors related to polarity. Best single model result in bold.

<table>
<thead>
<tr>
<th>system</th>
<th>negation insertion 1297</th>
<th>kein 10219</th>
<th>un- 11244</th>
<th>negation deletion 2919</th>
<th>kein 515</th>
<th>un- 586</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPE-to-BPE</td>
<td>94.8</td>
<td>99.1</td>
<td>97.1</td>
<td>93.8</td>
<td>88.7</td>
<td>86.5</td>
</tr>
<tr>
<td>BPE-to-char</td>
<td>92.7</td>
<td>98.9</td>
<td>98.7</td>
<td>91.0</td>
<td>85.1</td>
<td>78.8</td>
</tr>
<tr>
<td>char-to-char</td>
<td>92.1</td>
<td>98.9</td>
<td>98.8</td>
<td>91.5</td>
<td>86.4</td>
<td>80.5</td>
</tr>
<tr>
<td>(Sennrich et al., 2016a)</td>
<td>97.1</td>
<td>99.7</td>
<td>98.0</td>
<td>93.8</td>
<td>92.0</td>
<td>88.4</td>
</tr>
</tbody>
</table>

3.2 Results

Firstly, we report case-sensitive BLEU scores for all systems we trained for comparison to previous work.\(^{10}\) Results are shown in Table 3. The results confirm that our systems are comparable to previously reported results (Sennrich et al., 2016a; Chung et al., 2016), and that performance of the three systems is relatively close in terms of BLEU. The metric does not provide any insight into the respective strengths and weaknesses of different text representations.

Our main result is the assessment via contrastive translation pairs, shown in Table 4. We find that despite obtaining similar BLEU scores, the models have learned different structures to a different degree. The models with character decoder make fewer transliteration errors than the BPE-to-BPE model. However, they perform more poorly on separable verb particles and agreement, especially as distance increases, as seen in Figure 1. While accuracy for subject-verb agreement of adjacent words is similar across systems (95.2%, 94.0%, and 94.5% for BPE-to-BPE, BPE-to-char, and char-to-char, respectively), the gap widens for agreement between distant words – for a distance of over 15 words, the accuracy is 90.7%, 85.2%, and 82.3%, respectively.

Polarity shifts between the source and target text are a well-known translation problem, and our analysis shows that the main type of error is the deletion of negation markers, in line with findings of previous studies (Fancellu and Weber, 2015). We consider the relatively high number of errors related to polarity an important problem in machine translation, and hope that future work will try to improve upon our results, shown in more detail in Table 5.

\(^{10}\)Two commonly used BLEU evaluation scripts, the NIST BLEU scorer mteval-v13a.pl on detokenized text, and multi-bleu.perl on tokenized text, give different results due to tokenization differences. We here report both for comparison, but encourage the use of the NIST scorer, which is used by the WMT and IWSLT shared tasks, and allows for comparison of systems with different tokenizations.

We conclude from our results that there is currently a trade-off between generalization to unseen words, for which character-level decoders perform best, and sentence-level grammaticality, for which we observe better results with larger subword units of the BPE segmentation. We hope that our test set will help in developing and assessing architectures
Table 6: Examples where char-to-char model prefers contrastive translation (subject-verb agreement errors). 1-best translation can make error of same type (example 1), different type (translation of taught is missing in example 2), or no error (example 3).

<table>
<thead>
<tr>
<th>system</th>
<th>sentence</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>Since then we have only played in the Swedish league which is not the same level.</td>
<td>0.149</td>
</tr>
<tr>
<td>reference</td>
<td>Seitdem haben wir nur in der Schwedischen Liga gespielt, die nicht das gleiche Niveau hat.</td>
<td>0.137</td>
</tr>
<tr>
<td>contrastive</td>
<td>Seitdem haben wir nur in der schwedischen Liga gespielt, die nicht das gleiche Niveau haben.</td>
<td>0.090</td>
</tr>
<tr>
<td>1-best</td>
<td>Seitdem haben wir nur in der schwedischen Liga gespielt, die nicht die gleiche Stufe sind.</td>
<td>0.149</td>
</tr>
<tr>
<td>source</td>
<td>FriendsFest: the comedy show that taught us serious lessons about male friendship.</td>
<td>0.046</td>
</tr>
<tr>
<td>reference</td>
<td>FriendsFest: die Comedy-Show, die uns ernsthafte Lektionen über Männerfreundschaften erteilt hatte Robert Lewandowski.</td>
<td>0.276</td>
</tr>
<tr>
<td>contrastive</td>
<td>FriendsFest: die Comedy-Show, die uns ernsthafte Lektionen über Männerfreundschaften erteilen</td>
<td>0.262</td>
</tr>
<tr>
<td>1-best</td>
<td>FriendsFest: die Komödie zeigt, dass uns ernsthafte Lehren aus männlichen Freundschaften</td>
<td>0.129</td>
</tr>
</tbody>
</table>

We encourage the use of contrastive translation pairs, and LingEval97, for future analysis, but here discuss some limitations. The first one is by design: being focused on specific translation errors, the evaluation is not suitable as a global quality metric. Also, the evaluation only compares the probability of two translations, a reference translation $T$ and a contrastive translation $T'$, and makes no statement about the most probable translation $T^\ast$. Even if a model correctly estimates that $p(T) > p(T')$, it is possible that $T^\ast$ will contain an error of the same type as $T'$. And even if a model incorrectly estimates that $p(T) < p(T')$, it may produce a correct translation $T^\ast$. Despite these limitations, we argue that contrastive translation pairs are useful because they can easily be created to analyse any type of error in a way that is model-agnostic, automatic and reproducible.

Table 6 shows different examples of the where the contrastive translation is scored higher than the reference by the char-to-char model, and the corresponding 1-best translation. In the first one, our method automatically recognizes an error that also appears in the 1-best translation. In the second example, the 1-best translation is missing the verb. Such cases could confound a human analysis of agreement errors, and we consider it an advantage of our method that it is not confounded by other errors in the 1-best translation. In the third example, our method identifies an error, but the 1-best translation is correct. We note that the German reference exhibits object fronting, but the 1-best output has the more common SVO word order. While one could consider this instance a false positive, it can be important for an NMT model to properly score translations other than the 1-best, for instance for applications such as prefix-constrained MT (Wuebker et al., 2016).

4 Conclusion

We present LingEval97, a test set of 97 000 contrastive translation pairs for the assessment of neural machine translation systems. By introducing specific translation errors to the contrastive translations, we gain valuable insight into the ability of state-of-the-art neural MT systems to handle several challenging linguistic phenomena. A core finding is that recently proposed character-level decoders for neural machine translation outperform subword models at processing unknown names, but perform worse at modelling morphosyntactic agreement, where information needs to be carried over long distances. We encourage the use of LingEval97 to assess alternative architectures, such as hybrid word-character models (Luong and Manning, 2016), or dilated convolutional networks (Kalchbrenner et al., 2016). For the tested systems, the most challenging error type is the deletion of negation markers, and we hope that our test set will facilitate development and evaluation of models that try to improve in that respect. Finally, the evaluation via contrastive translation pairs is a very flexible approach, and can be applied to new language pairs and error types.

Acknowledgments

This project received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreements 645452 (QT21) and 688139 (SUMMA).
References


