

How Grammatical is Character-level Neural Machine Translation?

Assessing MT Quality with Contrastive Translation Pairs

Rico Sennrich

School of Informatics, University of Edinburgh
 {rico.sennrich}@ed.ac.uk

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 LingEval97¹, 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 a number of baseline results, and find that recently introduced character-level NMT systems perform better at transliteration than models with BPE segmentation, but perform more poorly at morphosyntactic agreement, and translating discontinuous 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 ques-

tion 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 BPE subword units (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 correct 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 phenomena that have previously been found to be challenging for machine translation, including agreement

¹Test set and evaluation script are available at <https://github.com/rsennrich/lingeal97>

over long distances (Koehn and Hoang, 2007; Williams and Koehn, 2011), discontinuous 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 task². Each contrastive translation pair consists of the correct reference translation, and a contrastive example that has been minimally modified to introduce a translation error. We define the accuracy of a model as the number of times it assigns a higher score to the correct 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 to 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 discontinuous 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 nega-

tion 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. 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 provide the less frequent one)

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 plausible errors. We expect that parsing errors will not invalidate the contrastive examples – correctly identifying the subject will affect the distance annotation, but changing the number of the verb should always introduce an error.³ Still, we report ceiling scores achievable by humans to account for the possibility that a generated error is not actually an error. We estimate the human ceiling by trying to select the correct variant for 20 contrastive translation pairs per category where our best system fails. The ceiling is below 100% because of errors in the reference translation, and cases that were undecidable by a human annotator (such as the gender of *the 20-year-old*).⁴

From the 22 191 sentences in the original newstest20** sets, we create approximately 97 000 contrastive translation pairs.

³Because of syncretism in German, there are cases where changing the inflection of one word does not cause disfluency, but merely changes the meaning. While a language model may deem both variants correct, a translation model should prefer the translation with the correct meaning.

⁴We mark all undecidable cases as wrong, and could perform better with random guessing.

²<http://www.statmt.org/wmt16/>

category	English	German (correct)	German (contrastive)
NP agreement	[...] of the American Congress	[...] des amerikanischen Kongresses	* [...] der amerikanischen Kongresses
subject-verb agr.	[...] that the plan will be approved	[...], dass der Plan verabschiedet wird	* [...], dass der Plan verabschiedet werden
separable verb particle	he is resting	er ruht sich aus	* er ruht sich an
polarity	the timing [...] is uncertain	das Timing [...] ist unsicher	das Timing [...] ist sicher
transliteration	Mr. Ensign's office	Senator Ensigns Büro	Senator Enisngs Büro

Table 1: Example contrastive translations pair for each error category.

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 produce unseen words?
- sequence-length is increased in character-level models, compared to word-level or BPE-level models. Does this have a negative effect on coherence or grammaticality?

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, Nematus⁵ for BPE-to-BPE, and dl4mt-c2c⁶ 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 tutorial⁷, 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

	BPE-BPE	BPE-char	char-char
source vocab	83,227	24,440	304
target vocab	91,000	302	302
source emb.	512	512	128
source conv.	-	-	(Lee et al., 2016)
target emb.	512	512	512
encoder	gru	gru	gru
encoder size	1024	512	512
decoder	gru_cond	two_layer_gru_decoder	
decoder size	1024	1024	1024
minibatch size	128	128	64
optimizer	adam	adam	adam
learning rate	0.0001	0.0001	0.0001
beam size	12	20	20
training time (minibatches)	≈ 1 week 240,000	≈ 2 weeks 510,000	≈ 2 weeks 540,000

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

system (test set and size→)	2014 3003	2015 2169	2016 2999
BPE-to-BPE	20.1 (21.0)	23.2 (23.0)	26.7 (26.5)
BPE-to-char	19.4 (20.5)	22.7 (22.6)	26.0 (25.9)
char-to-char	19.7 (20.7)	22.9 (22.7)	26.2 (26.1)
(Sennrich et al., 2016a)	25.4 (26.5)	28.1 (28.3)	34.2 (34.2)

Table 3: Case-sensitive BLEU scores (EN-DE) on WMT newstest. We report scores with detokenized NIST BLEU (mteval-v13a.pl), and in brackets, tokenized BLEU with multi-bleu.perl.

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 online⁸. It is also a BPE-to-BPE system, but in contrast to the previous systems, it includes different preprocessing (including truecasing), other hyperparameters, additional monolingual training data, an ensemble of models, and bidirectional decoding.

3.2 Results

Firstly, we report case-sensitive BLEU scores for all systems we trained for comparison to previous work.⁹ Results are shown in Table 3. The results

⁵<https://github.com/rsennrich/nematus>

⁶<https://github.com/nyu-dl/dl4mt-c2c>

⁷<https://github.com/nyu-dl/dl4mt-tutorial>

⁸http://data.statmt.org/rsennrich/wmt16_systems/

⁹Two commonly used BLEU evaluation scripts, the NIST BLEU scorer mteval-v13a.pl on detokenized text, and

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

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).

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.

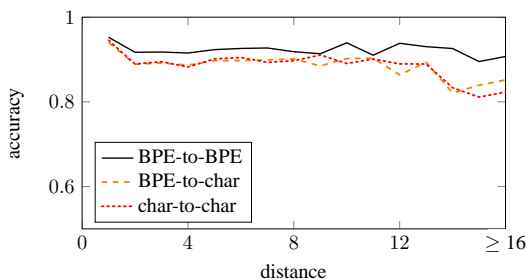


Figure 1: Subject-verb agreement accuracy as a function of distance between subject and verb.

Polarity shifts between the source and target text are a well-known translation problem, and our

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.

system (category and set size→)	negation insertion			negation deletion		
	nicht 1297	kein 10219	un- 11244	nicht 2919	kein 538	un- 586
BPE-to-BPE	94.8	99.1	97.1	93.0	88.7	86.5
BPE-to-char	92.7	98.9	98.7	91.0	85.1	78.8
char-to-char	92.1	98.9	98.8	91.5	86.4	80.5
(Sennrich et al., 2016a)	97.1	99.7	98.0	93.6	92.0	88.4

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

analysis shows that the main type of error is the deletion of negation markers, in line with findings of previous studies (Fancellu and Webber, 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 baseline results, shown in more detail in Table 5.

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 that aim to overcome this trade-off and perform best in respect to both morphology and syntax.

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^* . Even if a model correctly estimates that $p(T) > p(T')$, it is possible that T^* 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^* . 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.

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 words, 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 our baseline 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

- [Bahdanau et al.2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- [Chen and Zhu2014] Boxing Chen and Xiaodan Zhu. 2014. Bilingual Sentiment Consistency for Statistical Machine Translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 607–615, Gothenburg, Sweden, April. Association for Computational Linguistics.
- [Chung et al.2016] Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A Character-level Decoder without Explicit Segmentation for Neural Machine Translation. *CoRR*, abs/1603.06147.
- [Durrani et al.2014] Nadir Durrani, Hassan Sajjad, Hieu Hoang, and Philipp Koehn. 2014. Integrating an Unsupervised Transliteration Model into Statistical Machine Translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*, pages 148–153, Gothenburg, Sweden.
- [Fancellu and Webber2015] Federico Fancellu and Bonnie Webber. 2015. Translating Negation: A Manual Error Analysis. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 2–11, Denver, Colorado. Association for Computational Linguistics.
- [Kalchbrenner and Blunsom2013] Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle. Association for Computational Linguistics.
- [Kalchbrenner et al.2016] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural Machine Translation in Linear Time. *ArXiv e-prints*.
- [Koehn and Hoang2007] Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic. Association for Computational Linguistics.
- [Lee et al.2016] Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *ArXiv e-prints*, October.
- [Linzen et al.2016] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *ArXiv e-prints*, November.
- [Loáiciga and Gulordava2016] Sharid Loáiciga and Kristina Gulordava. 2016. Discontinuous Verb Phrases in Parsing and Machine Translation of English and German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.
- [Luong and Manning2016] Minh-Thang Luong and D. Christopher Manning. 2016. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063. Association for Computational Linguistics.

- [Nießen and Ney2000] Sonja Nießen and Hermann Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *18th Int. Conf. on Computational Linguistics*, pages 1081–1085.
- [Papineni et al.2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, PA. Association for Computational Linguistics.
- [Popovic2011] Maja Popovic. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *Prague Bull. Math. Linguistics*, 96:59–68.
- [Schmid et al.2004] Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. A German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1263–1266.
- [Sennrich and Haddow2015] Rico Sennrich and Barry Haddow. 2015. A Joint Dependency Model of Morphological and Syntactic Structure for Statistical Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2081–2087, Lisbon, Portugal. Association for Computational Linguistics.
- [Sennrich and Kunz2014] Rico Sennrich and Beat Kunz. 2014. Zmorge: A German Morphological Lexicon Extracted from Wiktionary. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.
- [Sennrich et al.2013] Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2013*, pages 601–609, Hissar, Bulgaria.
- [Sennrich et al.2016a] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 368–373, Berlin, Germany, August. Association for Computational Linguistics.
- [Sennrich et al.2016b] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- [Sutskever et al.2014] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 3104–3112, Montreal, Quebec, Canada.
- [Wetzel and Bond2012] Dominikus Wetzel and Francis Bond. 2012. Enriching Parallel Corpora for Statistical Machine Translation with Semantic Negation Rephrasing. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 20–29, Jeju, Republic of Korea, July. Association for Computational Linguistics.
- [Williams and Koehn2011] Philip Williams and Philipp Koehn. 2011. Agreement Constraints for Statistical Machine Translation into German. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 217–226, Edinburgh, UK. Association for Computational Linguistics.
- [Zeman et al.2011] Daniel Zeman, Mark Fishel, Jan Berka, and Ondřej Bojar. 2011. Addicter: What is wrong with my translations? *The Prague Bulletin of Mathematical Linguistics*, 96:79–88.