CharacTER: Translation Edit Rate on Character Level

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, Hermann Ney
Human Language Technology and Pattern Recognition, Computer Science Department
RWTH Aachen University, 52056 Aachen, Germany
<surname>@i6.informatik.rwth-aachen.de

Abstract

Recently, the capability of character-level evaluation measures for machine translation output has been confirmed by several metrics. This work proposes translation edit rate on character level (CharacTER), which calculates the character level edit distance while performing the shift edit on word level. The novel metric shows high system-level correlation with human rankings, especially for morphologically rich languages. It outperforms the strong CHRF by up to 7% correlation on different metric tasks. In addition, we apply the hypothesis sentence length for normalizing the edit distance in CharacTER, which also provides significant improvements compared to using the reference sentence length.

1 Introduction

The approaches for automatic evaluation of machine translation facilitated the development of statistical machine translation. They provide objective evaluation criteria for the translation results, and avoid the tedious and expensive manual evaluation. Currently the most commonly applied evaluation measures are the *Bilingual Evaluation Understudy* (BLEU) (Papineni et al., 2002) and the *Translation Edit Rate* (TER) (Snover et al., 2006) evaluation indicators. Most of the researchers use BLEU and TER as the primary metrics for evaluating their translation hypotheses.

The aim of the machine translation evaluation is to properly and objectively reflect the achievements and the functionality of machine translation. Through the evaluation, the developers of machine translation systems can learn the problems of the system and keep improving them. The evaluation metric not only provides the most reliable basis for machine translation systems, but also can be applied as the optimizing criterion in the parameter tuning step like BLEU. Thus, a good evaluation metric should demonstrate accuracy, universality and applicability.

In order to evaluate the applicability of different evaluation metrics, the correlation with human judgement is calculated. Currently the most common techniques for calculating the correlation between human and automatic evaluations are the *Spearman's rank correlation coefficient* (Spearman, 1904) and the *Pearson product-moment correlation coefficient* (Pearson, 1895).

In the recent past, several groups have reported further evaluation metrics, such as BEER (Stanojević and Sima'an, 2014) and CHRF (Popović, 2015), which actually outperformed the classic BLEU and TER metrics on Spearman and Pearson correlation with human judgement. In this work, we propose a novel *translation edit rate on character level* (CharacTER), which achieves a better correlation on the system-level for four different morphologically rich languages compared to BEER and CHRF. In addition, we also found that if we apply the hypothesis sentence length instead of reference sentence length to normalize the edit distance, the correlations of TER and CharacTER are improved by up to 9% on different languages.

2 Related Work

The most related work is the widely applied TER metric (Snover et al., 2006), which evaluates the amount of necessary editing to adjust a hypothesis so that it is accurately equal to a reference translation. Compared to the *Word Error Rate* (WER), TER introduced a *shift* edit on top of the *Levenshtein distance*, since in many languages different sequence orders are allowed. A hypothesis with

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 505–510,
Berlin, Germany, August 11-12, 2016.

another sequence order is not necessarily a bad translation. The TER is calculated by normalizing the total cost of edits over the entire sentence. The CharacTER inherits the word-level shift technique applied in TER and splits the shifted words into characters to calculate the edit distance.

This work is mainly motivated by (Popović, 2015), who proposed to apply character n-grams for automatic machine translation evaluation and achieved promising correlations. In this work, we will demonstrate that the TER on character level can also show a good performance, especially for morphologically rich languages, in which TER may miss matches due to various suffixes.

In addition to TER and CHRF, several other works are dedicated to the measurement of lexical similarity. These include, the commonly applied BLEU metric (Papineni et al., 2002), which calculates the geometric mean of the n-gram precision in a hypothesis based on a reference, and the METEOR metric (Lavie and Agarwal, 2007), which computes unigram overlaps between hypothesis and reference sequences considering stem matches and synonyms. The correlation of further evaluation metrics such as NIST (Doddington, 2002) and BEER (Stanojević and Sima'an, 2014) with human judgement are also presented in Section 4.

3 Character Level Edit Rate

Similar to TER, CharacTER is specified as the minimum number of character edits required to adjust a hypothesis, so that it absolutely matches the reference, normalized by the length of the hypothesis sentence (Equation 1). Note that here we apply the hypothesis instead of reference sentence length for normalization.

$$\text{CharacTER} = \frac{\text{shift cost} + \text{edit distance}}{\#\text{characters in the hypothesis sentence}} \quad (1)$$

3.1 Shift Edit

Unlike in speech or handwriting recognition, the *Character Error Rate* (CER) was not widely applied in machine translation. That is mainly because the shift edit is introduced for the translation metric, which is not necessary in speech recognition. In the calculation of TER, a greedy search is applied to discover the batch of shifts, by picking out the shift which most decreases the edit distance over and over again, until no more advantageous shifts exist. In other words, the shift edit is based on searching matched phrases between hypothesis and reference. Since the alphabet size in each language is very limited compared to the vocabulary size, characters are more likely to match each other than words, and thus the shift edit on the character level may corrupt words into meaningless pieces (Figure 1).

Besides the misplacement, the computational time is another big issue for directly applying TER on the character level. On the word level, we go through the hypothesis and compare the current word with each word in the reference. If a matched word is found at a different position in the reference, the succeeding words of the current word will be iteratively compared, in order to discover the longest matched phrases. This procedure becomes expensive on character level. The much higher matching probability of characters compared to words will result in many computations. For instance, for the example sentence in Figure 1, the computational time of the CER is 44 times as much as that of the TER.

In order to counter these issues, we applied a heuristic to calculate the translation edit rate on character level. Instead of shifting characters, we adopt the shift edit on word level as for calculating the TER. Then the shifted hypothesis sequence is split into characters and the Levenshtein distance is computed. In this way the computational time only increases by about 10% in our experiments. Note that here we consider the spaces in each sentence as extra characters, unlike in CHRF, since the correlation scores (Table 1) confirm the utility of this variant. We applied two different shift or phrase matching criteria: Two words are considered to be matched if

1. they are exactly the same,
2. or the edit distance between them is below a threshold value.

The first variant is the same as the phrase matching criterion in TER. In the second variant, the aim of introducing the threshold is to capture word pairs with the same stem, like `code` and `codes`. For the example in Figure 1, if we set a distance threshold to be 1, the shifted hypothesis sentence will be:

```
saudis the denied this week
information published in the new
your times
```

where `saudis` and `the` changed their positions

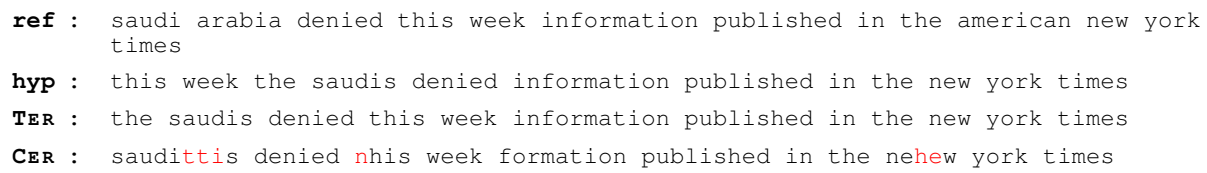

Figure 1: The hypothesis sentence after shift edit according to TER and CER technique. The characters marked with red color are the ones which are misplaced by the character level shift edit.

resulting in a smaller edit distance in this case. Based on the fact that the tolerance should be the same for long and short words, we applied absolute distances instead of ratios. For instance, if we use an error rate of 0.2 as the threshold value, words `eat` and `eats` are not considered to be matched, while words `translation` and `transition` will be matched. This issue can be fixed if we use an absolute edit distance equal to 1 as the threshold.

Another variant is the shift cost. In the calculation of TER, the shift of one entire phrase has a cost of 1, no matter how far this phrase moves. This penalty would be too mild for CharacTER, since the costs of insertions, deletions and substitutions become much larger on character level. Thus, we apply the average word length of the shifted phrase as the cost. For instance, the shift cost of phrase `the day before yesterday` will be $\frac{3+3+6+9}{4} = 5.25$. We also tried other possible costs, such as a fixed value or average word length of the whole data set. The experimental results are shown in Section 4.1.

3.2 Normalization

Both WER and TER techniques utilize a normalization over the reference sentence length by default, because the length of reference sentences stays unchanged, while different systems provide different translations with different hypothesis sentence lengths. In this case, the same edit distance of two hypotheses to the reference also indicates that they have the same TER, and the length of different translations is not taken into consideration. In this work we take advantage of other normalization alternatives.

First we used the hypothesis sentence length for the normalization. That means, with the same edit distance, the longer hypothesis results in a smaller error rate. For instance, let us consider the following reference and corresponding hypothesis sentences:

ref : `this is actually an estimate`
hyp$_1$: `this is in fact an estimate`
hyp$_2$: `indeed this is an estimate`

Compared to the reference sentence, the edit distances of both hyp$_1$ and hyp$_2$ are 2. Normalizing over the reference length results in TER $= \frac{2}{5} = 0.4$ for both hypotheses, whereas using the hypothesis length provides different results for them, equal to 0.33 and 0.4 respectively.

We also used other normalizer, such as the average, maximum or minimum length of reference and hypothesis sentence. We also calculated a CharacTER based on the entire data set, for which we sum up the edit distances of all sentences and also normalize the sum over the number of characters in the whole data set. According to our experimental results (Table 1) of the human correlation scores, normalizing using hypothesis length outperforms the other options, which is the case in all conducted experiments for both TER and our CharacTER. We suppose that human prefers the longer one, if two translations have equal quality. In addition, we note that in our translation experiments the default TER setup is heavily influenced by the hypothesis length: With the same BLEU score, a shorter translation normally achieves lower TER. The normalization over the hypothesis sentence length can effectively counter this issue.

4 Experiments

The evaluation metrics are correlated with human rankings by means of Spearman's rank correlation coefficient for the WMT13 task (Macháček and Bojar, 2013) and Pearson product-moment correlation coefficient for the WMT14 task (Macháček and Bojar, 2014) and WMT15 task (Stanojević et al., 2015) on the system level. Through the experiments we aim to investigate the following points:

- What is the most suitable threshold value to

identify the phrase matching?

- What shift cost should we apply?
- Which normalizer performs better?
- How does CharacTER perform compared to other metrics?

4.1 Comparison of different variants

First of all we would like to find out which is the best variant of the CharacTER. We conduct experiments on different shift costs and normalizers as described in Section 3, the correlation scores on different metric tasks are shown in Table 1. `basic setup` indicates the default setup of our metric, namely using the average length of the shifted words as the shift cost, considering only the exactly same words or phrases as matching and normalizing by length of each reference sentence. Other variants have the following meaning:

`w/o space` leaving out spaces in sentences

`threshold` the threshold edit distance to identify word matching

`shift` the shift cost of a phrase

`average` normalization over the average length of hypothesis and reference sentences

`max` normalization over the maximum length of hypothesis and reference sentences

`hyp` normalization over length of the hypothesis sentence

`whole` sum and normalization at the data set level instead of the sentence level

We also conducted experiments on other variants and variant combinations, such as other threshold values or shift costs. Only the variants with relative high correlation are presented in Table 1.

First of all, using the hypothesis sentence length as normalizer provides considerable improvements for both CharacTER and TER. Thus, we initiate to apply the hypothesis sentence length for normalizing not only our CharacTER but also the widely-used TER. Besides that, using an edit distance threshold also achieves significant improvements, while other configuration variants do not seem to be helpful. Thus on the following demonstrated experiments as well as on the shared metric task 2016, the configuration of CharacTER is organized as follows (the row with a cyan background in Table 1):

	WMT13		WMT14	
	en-*	*-en	en-*	*-en
TER	0.824	0.805	0.795	0.852
+ hyp	0.842	0.894	0.860	0.853
basic setup	0.857	0.832	0.833	0.868
+ w/o space	0.837	0.796	0.833	0.847
+ threshold 1	0.880	0.839	0.882	0.876
+ threshold 2	0.867	0.822	0.865	0.855
+ shift 1	0.836	0.813	0.820	0.847
+ shift 3	0.849	0.824	0.830	0.860
+ shift 5	0.839	0.818	0.836	0.866
+ average	0.917	0.913	0.871	0.928
+ max	0.908	0.918	0.849	0.918
+ hyp	0.925	0.928	0.908	0.930
+ whole	0.927	**0.931**	0.896	0.916
+ threshold 1	**0.934**	0.928	**0.916**	**0.938**

Table 1: Average correlations on WMT13 (Spearman) and WMT14 (Pearson) tasks for different variants of CharacTER. `en-*` indicates the average correlation for translations out of English, while `*-en` the translations into English. The best results in each direction are in bold.

- threshold value 1 to identify word matching
- average length of shifted words as shift cost
- hypothesis sentence length for normalization
- spaces in each sentence as extra characters

4.2 Comparison with other metrics

In this part the comparisons among different evaluation metrics are conducted. The correlations on different language pairs for the CharacTER metric along with the three mostly applied metrics BLEU, TER and METEOR, as well as the well-performing metrics for the corresponding tasks, are demonstrated in Table 2. The CharacTER metric performs quite well for out of English direction, especially on English→Russian, English→German and English→French tasks. On average we get up to 7% improvement compared to other strong metrics. It is noteworthy that on the WMT14 English→German task the CharacTER still provides a strong correlation, while other automatic metrics are negatively influenced by a large number of engaged systems of comparable quality. Additionally we list the best performing metrics in the WMT16 metrics task (Bojar et al., 2016) in Table 3. CharacTER surpasses other strong

WMT13	en-fr	en-de	en-es	en-cs	en-ru	avg.	fr-en	de-en	es-en	cs-en	ru-en	avg.
CharacTER	**0.944**	**0.926**	**0.916**	0.926	**0.957**	**0.934**	0.966	0.952	0.953	0.938	0.830	0.928
CHRF3[1]	0.914	0.919	0.758	0.895	0.820	0.861	0.984	**0.980**	**0.986**	**0.991**	**0.889**	**0.966**
SIMPBLEU[2]	0.924	0.925	0.830	0.867	0.710	0.851	0.978	0.936	0.923	0.909	0.798	0.909
BLEU	0.917	0.832	0.764	0.895	0.657	0.813	**0.989**	0.902	0.895	0.936	0.695	0.883
TER	0.912	0.854	0.753	0.860	0.538	0.783	0.951	0.833	0.825	0.800	0.581	0.798
METEOR	0.924	0.879	0.780	**0.937**	0.569	0.818	0.984	0.961	0.979	0.964	0.789	0.935

WMT14	en-fr	en-hi	en-cs	en-ru	avg.*	en-de	fr-en	de-en	hi-en	cs-en	ru-en	avg.
CharacTER	**0.957**	0.965	0.974	**0.933**	**0.958**	**0.757**	**0.976**	**0.957**	0.927	**0.986**	**0.844**	**0.938**
CHRF3	0.937	0.976	0.978	0.919	0.952	0.425	0.971	0.938	0.597	0.974	0.816	0.859
NIST[3]	0.941	**0.981**	**0.985**	0.927	**0.958**	0.200	0.955	0.811	0.784	0.983	0.800	0.867
BLEU	0.937	0.973	0.976	0.915	0.950	0.216	0.952	0.832	**0.956**	0.909	0.789	0.888
TER	0.954	0.829	0.978	0.931	0.923	0.324	0.952	0.775	0.618	0.976	0.809	0.826
METEOR	0.941	0.975	0.976	0.923	0.953	0.263	0.975	0.927	0.457	0.980	0.805	0.829

WMT15	en-fr	en-fi	en-de	en-cs	en-ru	avg.	fr-en	fi-en	de-en	cs-en	ru-en	avg.
CharacTER	0.942	0.854	**0.955**	0.970	**0.982**	**0.941**	**0.988**	0.888	**0.972**	0.960	0.884	0.939
CHRF3	0.932	**0.878**	0.848	**0.977**	0.946	0.916	0.979	0.903	0.956	0.968	0.898	0.941
BEER[4]	**0.961**	0.808	0.879	0.962	0.970	0.916	0.979	**0.965**	0.946	**0.983**	0.971	**0.969**
BLEU	0.948	0.602	0.573	0.936	0.841	0.780	0.975	0.929	0.865	0.957	0.851	0.915
TER	0.948	0.614	0.564	0.917	0.883	0.785	0.979	0.872	0.890	0.907	0.907	0.911
METEOR	0.959	0.760	0.650	0.953	0.892	0.843	0.982	0.950	0.953	**0.983**	**0.976**	**0.969**

Table 2: System-level correlations of automatic evaluation metrics and the official WMT human scores. The best results in each direction are in bold. We calculated the CharacTER and CHRF3 scores and cited the other scores from the WMT metric papers (Macháček and Bojar, 2013; Macháček and Bojar, 2014; Stanojević et al., 2015).

* English→German scores are not included in the averages of the WMT14 metrick task.

[1] CHRF3 (Popović, 2015)

[2] SIMPBLEU-RECALL (Song et al., 2013)

[3] NIST (Doddington, 2002)

[4] BEER (Stanojević and Sima'an, 2014)

WMT16	en-cs	en-de	en-fi	en-ro	en-ru	en-tr	cs-en	de-en	fi-en	ro-en	ru-en	tr-en
CharacTER	0.779	**0.915**	0.933	**0.959**	**0.954**	0.930	**0.997**	**0.985**	0.921	**0.970**	0.955	0.799
MPEDA	0.977	0.684	0.944	0.786	0.856	0.860	0.996	0.956	**0.967**	0.938	0.986	0.972
CHRF3	0.935	0.745	**0.974**	0.818	0.936	0.916	0.991	0.958	0.946	0.915	0.981	0.918
UOW.REVAL	-	-	-	-	-	-	0.993	0.949	0.958	0.919	**0.990**	0.977
BEER	0.972	0.732	0.940	0.947	0.906	**0.956**	0.996	0.949	0.964	0.908	0.986	**0.981**
WORDF3	**0.989**	0.768	0.901	0.931	0.836	0.714	0.991	0.898	0.786	0.909	0.955	0.803

Table 3: The preliminary results of the WMT16 metrics task: Absolute Pearson correlation of out-of-English and to-English system-level metric scores. All results are cited from (Bojar et al., 2016).

metrics on half of the language pairs. It performs especially well for English↔German and English↔Romanian. The results in Table 2 and 3 show that the CharacTER outperforms all other metrics on English→German by a large margin.

5 Conclusions

The experimental results showed in this paper exhibit that the translation edit rate on character level CharacTER represents a metric with high human correlations on the system-level, especially for the morphologically rich languages, which benefits from the character level information. We show the promising performance, while the concept is simple and straightforward. It is also noteworthy that the hypothesis sentence length is a better normalizer for both TER and CharacTER compared to the reference sentence length. As future work,

we would like to apply CharacTER as optimization criterion and conduct more experiments on non-European languages such as Chinese and Arabic.

Acknowledgments

This paper has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement nº 645452 (QT21).

References

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 Metrics Shared Task. In *Proceedings of the First Conference on Statistical Machine Translation*, Berlin, Germany, August.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 128–132, San Diego, CA, USA, March.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the ACL 2007 Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria, August.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, MA, USA, June.

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, USA, July.

Karl Pearson. 1895. Notes on Regression and Inheritance in the Case of Two Parents. In *Proceedings of the Royal Society of London*, volume 58, pages 240–242, London, UK, June.

Maja Popović. 2015. CHRF: Character n-gram F-Score for Automatic MT Evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisboa, Portugal, September.

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 the Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, USA, August.

Xingyi Song, Trevor Cohn, and Lucia Specia. 2013. BLEU deconstructed: Designing a better MT evaluation metric. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, Samos, Greece, March.

Charles Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15:72–101, January.

Miloš Stanojević and Khalil Sima'an. 2014. BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, MA, USA, June.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar, 2015. *Results of the WMT15 Metrics Shared Task*. Association for Computational Linguistics.