CHRF: character n-gram F-score for automatic MT evaluation

Maja Popović
Humboldt University of Berlin
Germany
maja.popovic@hu-berlin.de

Abstract

We propose the use of character n-gram F-score for automatic evaluation of machine translation output. Character n-grams have already been used as a part of more complex metrics, but their individual potential has not been investigated yet. We report system-level correlations with human rankings for 6-gram F1-score (CHRF) on the WMT12, WMT13 and WMT14 data as well as segment-level correlation for 6-gram F1 (CHRF) and F3-scores (CHRF3) on WMT14 data for all available target languages. The results are very promising, especially for the CHRF3 score – for translation from English, this variant showed the highest segment-level correlations outperforming even the best metrics on the WMT14 shared evaluation task.

1 Introduction

Recent investigations have shown that character level n-grams play an important role for automatic evaluation as a part of more complex metrics such as MTERATER (Parton et al., 2011) and BEER (Stanojević and Sima'an, 2014a; Stanojević and Sima'an, 2014b). However, they have not been investigated as an individual metric so far. On the other hand, the n-gram based F-scores, especially the linguistically motivated ones based on Part-of-Speech tags and morphemes (Popović, 2011), are shown to correlate very well with human judgments clearly outperforming the widely used metrics such as BLEU and TER.

In this work, we propose the use of the F-score based on character n-grams, i.e. the CHRF score. We believe that this score has a potential as a stand-alone metric because it is shown to be an important part of the previously mentioned complex measures, and because, similarly to the morpheme-based F-score, it takes into account some morpho-syntactic phenomena. Apart from that, in contrast to the related metrics, it is simple, it does not require any additional tools and/or knowledge sources, it is absolutely language independent and also tokenisation independent.

The CHRF scores were calculated for all available translation outputs from the WMT12 (Callison-Burch et al., 2012), WMT13 (Bojar et al., 2013) and WMT14 (Bojar et al., 2014) shared tasks, and then compared with human rankings. System-level correlation coefficients are calculated for all data, segment-level correlations only for WMT14 data. The scores were calculated for all available target languages, namely English, Spanish, French, German, Czech, Russian and Hindi.

2 CHRF score

The general formula for the CHRF score is:

$$\text{CHRF}\beta = (1+\beta^2)\frac{\text{CHRP}\cdot\text{CHRR}}{\beta^2\cdot\text{CHRP}+\text{CHRR}} \quad (1)$$

where CHRP and CHRR stand for character n-gram precision and recall arithmetically averaged over all n-grams:

- CHRP
  percentage of n-grams in the hypothesis which have a counterpart in the reference;

- CHRR
  percentage of character n-grams in the reference which are also present in the hypothesis.

and β is a parameter which assigns β times more importance to recall than to precision – if $\beta = 1$, they have the same importance.

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395,
Lisboa, Portugal, 17-18 September 2015.

3 Experiments on WMT12, WMT13 and WMT14 test data

3.1 Experiments

As a first step, we carried out several experiments regarding n-gram length. Since the optimal n for word-based measures is shown to be $n = 4$, MTERATER used up to 10-gram and BEER up to 6-gram, we investigated those three variants. In addition, we investigated a dynamic n calculated for each sentence as the average word length. The best correlations are obtained for 6-gram, therefore we carried out further experiments only on them.

Apart from the n-gram length, we investigated the influence of the space treating it as an additional character. However, taking space into account did not yield any improvement regarding the correlations and therefore has been abandoned.

words	This is an example.
characters	T h i s i s a n e x a m p l e .
+space	T h i s _ i s _ a n _ e x a m p l e .

Table 1: Example of an English sentence with its corresponding character sequences without and with taking the space into account.

In the last stage of our current experiments, we have compared two β values on the WMT14 data: the standard CHRF with $\beta = 1$ i.e. the harmonic mean of precision and recall, as well as CHRF3 where $\beta = 3$, i.e. the recall has three times more weight. The number 3 has been taken arbitraly as a preliminary value, and the CHRF3 is tested only on WMT14 data – more systematic experiments in this direction should be carried out in the future work.

3.2 Correlations with human rankings

System-level correlations

The evaluation metrics were compared with human rankings on the system-level by means of Spearman's correlation coefficients ρ for the WMT12 and WMT13 data and Pearson's correlation coefficients r for the WMT14 data. Spearman's rank correlation coefficient is equivalent to Pearson correlation on ranks, and it makes fewer assumptions about the data.

Average system-level correlations for CHRF score(s) together with the word n-gram F-score WORDF and the three mostly used metrics BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and METEOR (Banerjee and Lavie, 2005) are shown in Table 2. It can be seen that the CHRF score is comparable or better than the other metrics, especially the CHRF3 score.

Table 3 presents the percentage of translation outputs where the particular F-score metric (WORDF, CHRF and CHRF3) has higher correlation (no ties) than the particular standard metric (BLEU, TER and METEOR). It can be seen that the WORDF score outperforms BLEU and TER for about 60% of documents, however METEOR only in less than 40%. Standard CHRF is better than METEOR for half of the documents, and better than BLEU and TER for 68% of the documents thus being definitely more promising than the word-based metrics. Finally, CHRF3 score outperforms all standard metric for about 70-80% of texts, thus being the most promising variant.

Segment-level correlations

The segment-level quality of metrics is measured using Kendall's τ rank correlation coefficient. It measures the metric's ability to predict the results of the manual pairwise comparison of two systems. The τ coefficients were calulated only on the WMT14 data using the official WMT14 script, and the obtained WMT14 variant is reported for the WORDF score, both CHRF scores, as well as for the best ranked metrics in the shared evaluation task.

Table 4 shows the τ coefficients for translation into English (above) and for translation from English (below). For translation into English, it can be seen that the CHRF3 score is again the most promising F-score. Furthermore, it can be seen that the correlations for both CHRF scores are close to the two best ranked metrics (DISCOTKPARTY and BEER) and the METEOR metrics, which is very well ranked too. For translation from English, the CHRF3 score yields the highest average correlation, and the CHRF score is comparable with the best ranked BEER metric.

4 Conclusions

The results presented in this paper show that the character n-gram F-score CHRF represents a promising metric for automatic evaluation of machine translation output for several reasons: it is language-independent, tokenisation-independent and it shows good correlations with human judgments both on the system- as well as

year	WORDF	CHRF	CHRF3	BLEU	TER	METEOR
2014 (r)	0.810	0.805	0.857	0.845	0.814	0.822
2013 (ρ)	0.874	0.873	/	0.835	0.791	0.876
2012 (ρ)	0.659	0.696	/	0.671	0.682	0.690

Table 2: Average system-level correlations on WMT14 (Pearson's r), WMT13 and WMT12 data (Spearman's ρ) for word 4-gram F1 score, character 6-gram F1 score and character 6-gram F3 score together with the three mostly used metrics BLEU, TER and METEOR.

rank>	WORDF	CHRF	CHRF3
BLEU	64.3	67.9	80.0
TER	60.7	67.9	70.0
METEOR	39.3	50.0	70.0

Table 3: *rank*> for three F-scores (WORDF, CHRF and CHRF3) in comparison with three standard metrics (BLEU, TER and METEOR) – percentage of translation outputs where the given F-score metrics has higher correlation than the given standard metric.

Kendall's τ	fr-en	de-en	hi-en	cs-en	ru-en	avg.
WORDF	0.356	0.258	0.276	0.200	0.262	0.270
CHRF	0.402	0.318	0.395	0.253	0.320	0.338
CHRF3	0.391	0.332	0.394	0.278	0.322	0.343
DISCOTKPARTY	0.433	0.380	0.434	0.328	0.355	0.386
BEER	0.417	0.337	0.438	0.284	0.333	0.362
METEOR	0.406	0.334	0.420	0.282	0.329	0.354

Kendall's τ	en-fr	en-de	en-hi	en-cs	en-ru	avg.
WORDF	0.251	0.205	0.202	0.281	0.381	0.264
CHRF	0.296	0.247	0.253	0.331	0.443	0.314
CHRF3	0.304	0.269	0.294	0.331	0.457	0.331
BEER	0.292	0.268	0.250	0.344	0.440	0.319
METEOR	0.280	0.238	0.264	0.318	0.427	0.306

Table 4: Segment-level Kendall's τ correlations on WMT 14 data for WORDF, CHRF and CHRF3 score together with the best performing metrics on the shared evaluation task.

on the segment-level, especially the CHRF3 variant. Therefore both of the CHRF scores were submitted to the WMT15 shared metrics task. In future work, different β values should be investigated, as well as different weights for particular n-grams. Apart from this, CHRF is so far tested on only one non-European language (Hindi) – application on more languages using different writing systems such as Arabic, Chinese, etc. has to be explored systematically.

Acknowledgments

This publication has emanated from research supported by QTLEAP project (Quality Translation by Deep Language Engineering Approach) – ECs FP7 (FP7/2007-2013) under grant agreement number 610516, QT21 project funded by the European Union's Horizon 2020 research and innovation programme under grant number 645452, and TRAMOOC project (Translation for Massive Open Online Courses) partially funded by the European Commission under H2020-ICT-2014/H2020-ICT-2014-1 under grant agreement number 644333. Special thanks to Miloš Stanojević for suggesting experiments with the β parameter.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. In *Proceedings of the ACL 05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI, June.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT-13)*, pages 1–44, Sofia, Bulgaria, August.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT-14)*, page 1258, Baltimore, Maryland, USA, June.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation (WMT-12)*, page 1051, Montral, Canada, June. Association for Computational Linguistics.

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 of the Association for Computational Linguistics (ACL 02)*, pages 311–318, Philadelphia, PA, July.

Kristen Parton, Joel Tetreault, Nitin Madnani, and Martin Chodorow. 2011. E-Rating Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT-11)*, pages 108–115, Edinburgh, Scotland.

Maja Popović. 2011. Morphemes and POS tags for n-gram based evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT-11)*, pages 104–107, Edinburgh, Scotland, July.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 06)*, pages 223–231, Boston, MA, August.

Miloš Stanojević and Khalil Sima'an. 2014a. BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT-14)*, pages 414–419, Baltimore, Maryland, June.

Miloš Stanojević and Khalil Sima'an. 2014b. Fitting Sentence Level Translation Evaluation with Many Dense Features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-14)*, pages 202–206, Doha, Qatar, October. Association for Computational Linguistics.