CHRF ++: words helping character n-grams

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

Abstract

Character n-gram F-score (CHRF) is shown to correlate very well with human relative rankings of different machine translation outputs, especially for morphologically rich target languages. However, its relation with direct human assessments is not yet clear. In this work, Pearson's correlation coefficients for direct assessments are investigated for two currently available target languages, English and Russian. First, different β parameters (in range from 1 to 3) are re-investigated with direct assessment, and it is confirmed that $\beta = 2$ is the optimal option. Then separate character and word n-grams are investigated, and the main finding is that, apart from character n-grams, word 1-grams and 2-grams also correlate rather well with direct assessments. Further experiments show that adding word unigrams and bigrams to the standard CHRF score improves the correlations with direct assessments, though it is still not clear which option is better, unigrams only (CHRF+) or unigrams and bigrams (CHRF++). This should be investigated in future work on more target languages.

1 Introduction

Recent investigations (Popović, 2015; Stanojević et al., 2015; Popović, 2016; Bojar et al., 2016) have shown that the character n-gram F-score (CHRF) represents a very promising evaluation metric for machine translation, especially for morphologically rich target languages – it is fast, it does not require any additional tools or information, it is language independent and tokenisation independent, and it correlates very well with human relative rankings (RR) (Callison-Burch et al., 2008). In order to produce these rankings, human annotators have to decide which sentence translation is better/worse than another without giving any note about the absolute quality of any of the evaluated translations. This type of human judgment has been the offical evaluation metric and gold standard for all automatic metrics at WMT shared tasks from 2008 until 2016.

Another type of human judgment, direct human assessment (DA) (Bojar et al., 2016), has become additional official evaluation metric for WMT-16, and the only one for WMT-17. These assessments consist of absolute quality scores for each translated sentence. Contrary to RR, the relation between CHRF and DA has still not been investigated systematically. Preliminary experiments in previous work (Popović, 2016) shown that, concerning DA, the main advantage of character-based F-score CHRF in comparison to word-based F-score WORDF is better correlation for good translations for which WORDF often assigns too low scores.

In this work, we systematically investigate relations between DA and both character and word n-grams, as well as their combinations. The scores are calculated for all available translation outputs from the WMT-15 and WMT-16 shared tasks (Bojar et al., 2016) which contain two target languages, English (translated from Czech, German, Finnish, Romanian, Russian and Turkish) and Russian (translated from English), and then compared with DAs on segment level using Pearsons's correlation coefficient.

2 n-gram based F-scores

The general formula for an n-gram based F-score is:

$$ngrF\beta = (1 + \beta^2)\frac{ngrP \cdot ngrR}{\beta^2 \cdot ngrP + ngrR} \quad (1)$$

where $ngrP$ and $ngrR$ stand for n-gram precision and recall arithmetically averaged over all n-grams from n = 1 to N:

- $ngrP$
  n-gram precision: percentage of n-grams in the hypothesis which have a counterpart in the reference;

- $ngrR$
  n-gram recall: percentage of n-grams in the reference which are also present in the hypothesis.

and β is a parameter which assigns β times more weight to recall than to precision.

WORDF is then calculated on word n-grams and CHRF is calculated on character n-grams. As for maximum n-gram length N, previous work reported that there is no need to go beyond N=4 for WORDF (Popović, 2011) and N=6 for CHRF (Popović, 2015).

CHRF++ score is obtained when the word n-grams are added to the character n-grams and averaged together. The best maximum n-gram lengths for such combinations are again N=6 for character n-grams and N=2 or N=1 for word n-grams, which will be discussed in Section 4.3.

3 Motivation for adding word n-grams to CHRF

A preliminary experiment on a small set of texts reported in previous work (Popović, 2016) with different target languages and different types of DA[1] shown that for poorly rated sentences, the standard deviations of CHRF and WORDF scores are similar – both metrics assign relatively similar (low) scores. On the other hand, for the sentences with higher human rates, the deviations for CHRF are (much) lower. In addition, the higher the human rating is, the greater is the difference between the WORDF and CHRF deviations. These results indicate that CHRF is better than WORDF mainly for segments/systems of higher translation quality – the CHRF scores for good translations are more concentrated in the higher range, whereas the WORDF scores are often too low.

In order to further investigate these premises, scatter plots in Figure 1 are produced for CHRF and WORDF with DA for the Russian→English and English→Russian WMT-16 data.

[1] none of them equal to the variant used in WMT

Figure 1 confirms the findings from previous work, since a number of WORDF values is indeed pessimistic – high DA but low WORDF, whereas CHRF values are more concentrated, i.e. correlate better with DA values. However, these plots raised another question – are CHRF scores maybe too optimistic (i.e. segments with high CHRF score and low DA score)? Certainly not to such extent as WORDF scores are pessimistic, but still, could some combination of character and word n-grams improve the correlations of CHRF?

4 Pearson correlations with direct assessments

In order to explore combining CHRF with word n-grams, the following experiments are carried out in terms of calculating Pearson's correlation coefficient between DA and different n-gram F-scores:

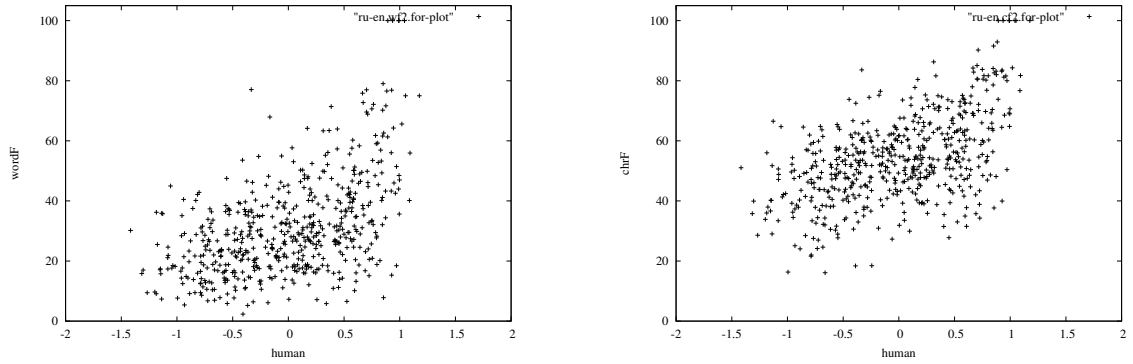
1. As a first step, β parameter is re-investigated for DA, both for CHRF and WORDF in order to check if β = 2 is a good option for DA, too;

2. Individual character and word n-grams are investigated in order to see if some are better than others and to which extent;

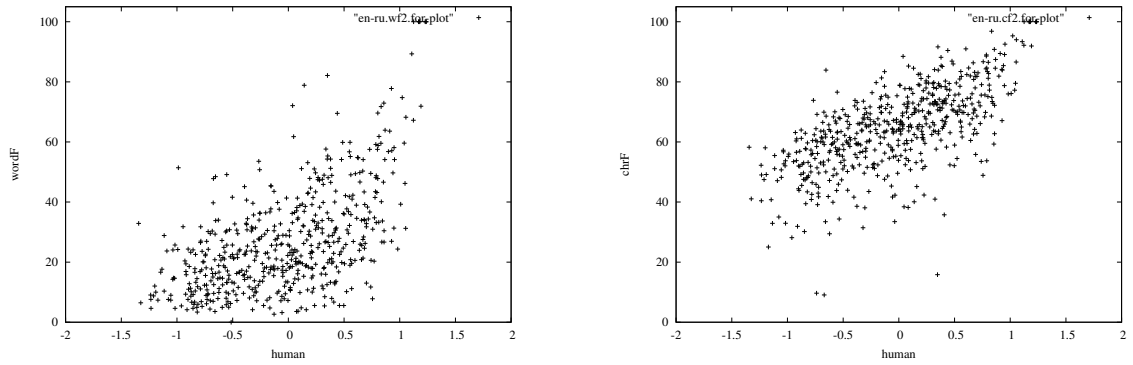
3. Finally, various combinations of character and word n-grams were explored and the results are reported for the most promising ones.

4.1 β parameter revisited

Previous work (Popović, 2016) reported that the best β parameter both for CHRF and for WORDF is 2 in terms of Kendall's τ segment level correlation with human relative rankings (RR). However, this parameter has not been tested for direct human assessments (DA) – therefore we tested several β in terms of Pearson correlations with DA. It is confirmed that putting more weight on precision is not good, and the results for β = 1,2,3 are reported in Table 1. Both for CHRF and WORDF, the correlations for β = 2,3 are comparable, and better than for β =1. Since there is almost no difference between 2 and 3, and putting too much weight to recall could jeopardise some other applications such as system tuning or system combination (for example, (Sánchez-Cartagena and Toral, 2016) decided to use CHRF1 because CHRF3 lead to generation of too long sentences), we decided to choose β = 2 which will be used for all further experiments.



(a) Russian→English, WORDF (c) Russian→English, CHRF



(b) English→Russian, WORDF (d) English→Russian, CHRF

Figure 1: Scatter plots for (a)(b) WORDF and (c)(d) CHRF with DA for (a)(c) Russian→English and (b)(d) English→Russian WMT-16 texts confirm that WORDF values are overly pessimistic – a number of WORDF points lies in the lower right quadrant, i.e. a number of segments with high DA values has a low WORDF value. On the other hand, CHRF points are more concentrated, especially for morphologically rich Russian. However, are some of them too optimistic? (i.e. segments with high CHRF scores and low DA scores)

4.2 Individual character and word n-grams

Individual n-grams were also investigated in previous work, however (i) only character n-grams and (ii) only compared with RR, not with DA. In this work, we carried out systematic investigation on both character and word n-grams' correlations with DA, and the results are reported in Table 2. It should be noted that, to the best of our knowledge, word n-grams with order less than 4 have not been investigated yet in the given context of correlations with RR or DA. Implicitly, the METEOR metric (Banerjee and Lavie, 2005) is based on word unigrams with additional information and generally correlates better with human rankings than the BLEU metric (Papineni et al., 2002) based on uni-, bi-, 3- and 4-gram precision.

The results show that, similarly to the correlations with RR, the best character n-grams are of the middle lengths i.e. 3 and 4. The main finding is, though, that the best word n-grams are the short ones, namely unigrams and bigrams.

Following these results for individual n-grams, several different experiments have been carried out, involving different character n-gram weights, combining character and word n-grams with different weights, etc., however no consistent improvements have been noticed in comparison to the standard uniform n-gram weights, not even by removing or setting low weight for character unigrams. The only noticeable improvement was observed when word 4-grams and 3-grams were removed.

4.3 The emergence of CHRF++

Findings reported in the previous section raised the following questions: (i) are word 3-grams and 4-grams the "culprits" for overly pessimistic behaviour of WORDF described in Section 3? (ii) Could the "good guys", i.e. word unigrams and

2016/2015	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	en-ru	*mean*
CHRF1	.644/.542	.452/.600	.454/.565	.570	.522/.601	.551	.642/.606	.562
CHRF2	.658/**.552**	.469/**.605**	**.457/.573**	.581	.534/.613	**.556**	**.661/.624**	**.574**
CHRF3	**.660/.552**	**.472**/.604	.455/.572	**.582**	**.535/.614**	.555	**.661**/.622	**.574**
WORDF1	.587/.503	.453/.540	.428/.525	.504	.498/.549	.531	.572/.527	.519
WORDF2	.598/.512	.462/.543	.437/.535	.518	.504/.559	.536	.580/.533	.526
WORDF3	.600/.514	.464/.543	.439/.537	.522	.504/.561	.536	.582/.534	.528

Table 1: Pearson's correlation coefficients of CHRF and WORDF with direct human assessments (DA) for different β parameters. Bold represents the best character level value and underline represents the best word level value. The best β values are 2 and 3.

2016/2015	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	en-ru	*mean*
chr1-gram	.544/.448	.355/.407	.313/.417	.443	.358/.527	.337	.531/.489	.431
chr2-gram	.644/.537	.441/.556	.420/.547	.554	.504/.599	.513	.652/.631	.550
chr3-gram	**.662**/.539	**.472**/.604	.459/**.582**	.579	.533/**.613**	.559	**.683/.661**	**.579**
chr4-gram	.657/**.542**	**.472/.614**	**.460**/.581	**.582**	**.538**/.602	**.562**	.682/.655	**.579**
chr5-gram	.644/.540	.467/.611	.456/.559	.576	.532/.588	.559	.676/.640	.571
chr6-gram	.627/.539	.463/.599	.447/.539	.568	.521/.578	.553	.662/.623	.560
word1-gram	.631/.509	.481/.529	.434/.566	.504	.505/.606	.510	.601/.564	.537
word2-gram	.611/.528	.473/.546	.441/.513	.529	.513/.551	.539	.575/.549	.531
word3-gram	.546/.461	.426/.513	.387/.470	.498	.469/.519	.475	.536/.472	.481
word4-gram	.479/.382	.385/.458	.337/.369	.427	.404/.468	.380	.478/.397	.414

Table 2: Pearson's correlation coefficients of CHRF and WORDF with direct human assessments (DA) for individual character and word n-grams. Bold represents the best character level value and underline represents the best word level value.

bigrams diminish potentially too optimistical behaviour of CHRF?

In order to get the answers, the Pearson correlations are calculated for CHRF combined with four WORDFs with different maximum n-gram lengths, i.e. N=1,2,3,4 and the results are presented in Table 3. In addition, correlations are presented also for CHRF and two variants of WORDF (usual N=4 and the best N=2).

First, it can be seen that removing word 3-grams and 4-grams improves the correlation for WORDF which becomes closer to CHRF (and even better for one of the two German→English texts). Furthermore, it can be seen that adding word unigrams and bigrams to CHRF improves the correlations of CHRF in the best way. Therefore this is the variant which is chosen to be the CHRF++. Next best option (CHRF+) is to add only word unigrams i.e. words, and this one is the best one for translation into Russian. Possible reasons are morphological richness of Russian as well as rather free word order, however the test set in this experiment is too small to draw any conclusions. Both CHRF++ and CHRF+ should be further tested on more texts and on more morphologically rich languages.

Scatter plots presented in Figure 2 visualise the improvement of correlations by CHRF++: WORDF with N=4 (a) is, as already shown, too pessimistic. Lowering the maximum n-gram length to 2 (b) moves a number of pessimistic points upwards, thus improving the correlation. When added to slightly overly optimistic CHRF (c), the points for both metrics are moved more towards the middle (d).

5 Conclusions

The results presented in this work show that adding short word n-grams, i.e. unigrams and bigrams to the character n-gram F-score CHRF improves the correlation with direct human assessments (DA). Since the amount of available texts with DA is still small, it is still not possible to conclude which variant is better: adding only unigrams (CHRF+) or unigrams and bigrams (CHRF++). This is especially hard to conclude for translation into morphologically rich languages, since only Russian was available until now. In order to explore both CHRF+ and CHRF++ more systematically, both are submitted to the WMT-17 metrics task for translations from English. For translation into English, only CHRF++ is submitted since it outperformed the other variant for English. For Chinese, only the raw CHRF has been submitted since the concept "Chinese words" is generally not clear. Further work should include more data and more distinct target languages.

The tool for calculating CHRF++ (as well as CHRF+ and CHRF since it is possible to change maximum n-gram lengths) is publicly available at `https://github.com/m-popovic/chrF`. It is a Python script which requires (multiple) reference translation(s) and a translation hypothesis (output) in the raw text format. It is language independent and does need tokenisation or any similar preprocessing of the text. The default β is set to 2, but it is possible to change. It provides both segment level scores as well as document level scores in two variants: micro- and macro-averaged.

Acknowledgments

This work has been supported by the TraMOOC project funded from the European Unions Horizon 2020 research and innovation programme under grant agreement No 644333.

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*. Ann Arbor, MI, pages 65–72.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 Metrics Shared Task. In *Proceedings of the First Conference on Machine Translation (WMT-16)*. Berlin, Germany, pages 199–231.

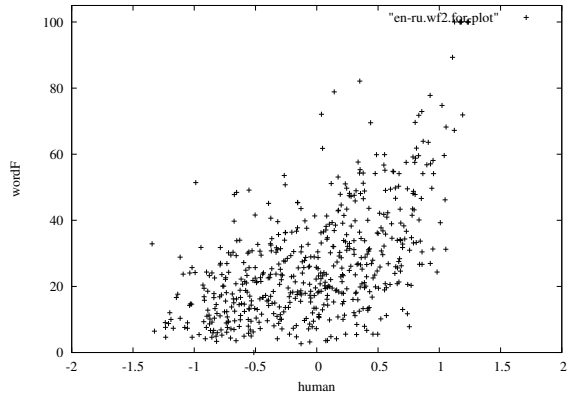
Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *Proceedings of the 3rd ACL 08 Workshop on Statistical Machine Translation (WMT-08)*. Columbus, Ohio, pages 70–106.

Kishore Papineni, Salim Roukos, Todd Ward, and Wie-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)*. Philadelphia, PA, pages 311–318.

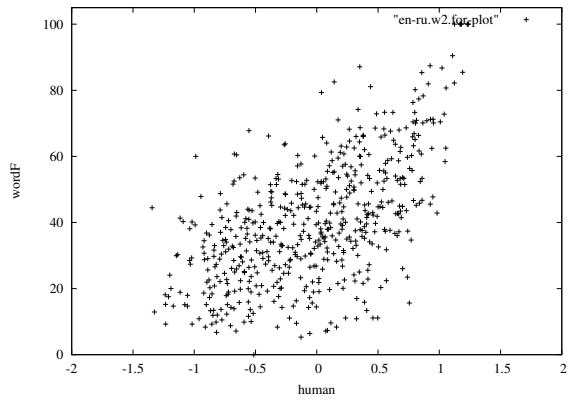
Maja Popović. 2011. Morphemes and POS tags for n-gram based evaluation metrics. In *Proceedings of*

2016/2015	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	en-ru	*mean*
WORDF (4-gram)	.598/.512	.462/.543	.437/.535	.518	.504/.559	.536	.580/.533	.526
WORD2-F	.642/.537	.490/.557	.455/.563	.535	.526/.592	.553	.603/.575	.552
CHRF (6-gram)	.658/.552	.469/.605	.457/.573	**.581**	.534/613	.556	.661/.624	.574
C6W4F	.656/.555	.483/.598	.470/.580	.572	.538/.608	.573	.665/.630	.577
C6W3F	.663/.559	.486/.603	**.471**/.584	.578	.542/.615	**.574**	.672/.641	.582
C6W2F (CHRF++)	**.668/.561**	**.487/.606**	.470/**.585**	.580	**.544**/.619	.570	.679/.650	**.585**
C6W1F (CHRF+)	.665/.558	.480/**.606**	.464/**.585**	.579	.540/**.620**	.562	**.685/.654**	.583

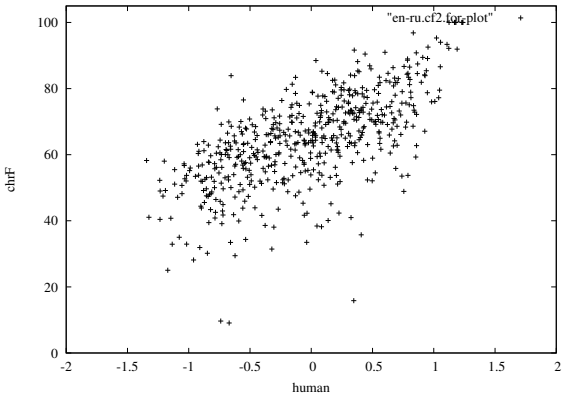
Table 3: Pearson's correlation coefficients with direct human assessments (DA) of CHRF enhanced with word n-grams together with CHRF and two variants of WORDF: N=4 and N=2. Bold represents the best overall value.



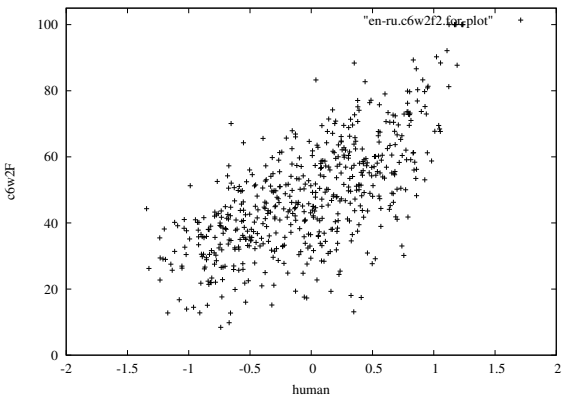
(a) wordF (n-gram order = 4)



(b) word2F (n-gram order = 2)



(c) CHRF (n-gram order = 6)



(d) CHRF++ (CHRF +word2F)

Figure 2: Scatter plots for (a) WORDF with N=4, (b) WORDF with N=2, (c) CHRF and (d) CHRF++ (CHRF enhanced with word bigrams) with DA for English→Russian WMT-16 text. Removing word 3-grams and 4-grams decreases the number of "pessimistic" WORDF points in the lower right quadrant. Combining CHRF with word unigrams and bigrams further decreases the frequency of such points and also lowers overall CHRF scores pushing the points more towards the middle.

the Sixth Workshop on Statistical Machine Translation (WMT-11). Edinburgh, Scotland, pages 104–107.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*. Lisbon, Portugal, pages 392–395.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation (WMT-16)*. Berlin, Germany, pages 499–504.

Víctor M. Sánchez-Cartagena and Antonio Toral. 2016. Abu-matran at wmt 2016 translation task: Deep learning, morphological segmentation and tuning on character sequences. In *Proceedings of the First Conference on Machine Translation (WMT-16)*. Berlin, Germany, pages 362–370.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 Metrics Shared Task. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*. Lisbon, Portugal, pages 256–273.