CHRF deconstructed: β parameters and n-gram weights

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

Abstract

Character n-gram F-score (CHRF) is shown to correlate very well with human rankings of different machine translation outputs, especially for morphologically rich target languages. However, only two versions have been explored so far, namely CHRF1 (standard F-score, β = 1) and CHRF3 (β = 3), both with uniform n-gram weights. In this work, we investigated CHRF in more details, namely β parameters in range from 1/6 to 6, and we found out that CHRF2 is the most promising version. Then we investigated different n-gram weights for CHRF2 and found out that the uniform weights are the best option. Apart from this, CHRF scores were systematically compared with WORDF scores, and a preliminary experiment carried out on small amount of data with direct human scores indicates that the main advantage of CHRF is that it does not penalise too hard acceptable variations in high quality translations.

1 Introduction

Recent investigations (Popović, 2015; Stanojević et al., 2015) 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 simple, it does not require any additional tools or information, it is language independent and tokenisation independent, and it correlates very well with human rankings. However, only two versions of this score have been investigated so far: standard F-score CHRF1 where β = 1, i.e. precision and recall have the same weight, as well as CHRF3, where recall has three times more weight.

In this work, we systematically investigate β parameters: standard version (β = 1), five β values favorising recall (2,3,4,5,6) and five β values favorising precision (1/2, 1/3, 1/4, 1/5 and 1/6). In addition, we also compare CHRFβ scores with WORDFβ scores.

The CHRFβ and WORDFβ scores are calculated for all available translation outputs from the WMT14 (Bojar et al., 2014) and WMT15 (Bojar et al., 2015) shared tasks and then compared with human rankings on segment level using Kendall's τ rank correlation coefficient.

The scores were analysed for all available target languages. i.e. English, French, German, Czech, Russian, Hindi and Finnish.

2 CHRF and WORDF scores

The general formula for 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. If $\beta = 1$, they have the same weight; if β = 4, recall has four times more importance than precision; if β = 1/4,

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 499–504,
Berlin, Germany, August 11-12, 2016.

precision has four times more importance than recall.
WORDF is then calculated on word n-grams and CHRF is calculated on character n-grams. Maximum n-gram length N for both metrics is investigated in previous work, and N=4 is shown to be optimal for WORDF (Popović, 2011), N=6 for CHRF (Popović, 2015).

3 Comparison of CHRFβ and WORDFβ scores

The CHRFβ and WORDFβ scores are calculated for the following β parameters: 1/6, 1/5, 1/4, 1/3, 1/2, 1, 2, 3, 4, 5 and 6. For each CHRFβ and WORDFβ score, the segment level τ correlation coefficients are calculated for each translation output. In total, 20 τ coefficients were obtained for each score – five English outputs from the WMT14 task and five from the WMT15, together with ten outputs in other languages, i.e. two French, two German, two Czech, two Russian, one Hindi and one Finnish. The obtained τ coefficients were then summarised into the following four values:

- *mean*
  τ averaged over all translation outputs;
- *diff*
  averaged difference between the τ of the particular metric and the τs of all other metrics investigated in this work;
- *rank*$>$
  percentage of translation outputs where the particular metric has better τ than the other metrics investigated in this work;
- *rank*$\geqslant$
  percentage of translation outputs where the particular metric has better or equal τ than the other metrics investigated in this work.

These values for each metric are presented in Table 1. In addition, the values are shown separately for translation into English (Table 2) and for translation out of English (Table 3).
Table 1 shows that:

- CHRF ranks better than WORDF;
- recall is more important than precision;
- the most promising metric is CHRF2;

metric	*mean*	*diff*	*rank*$>$	*rank*$\geqslant$
CHRF1/6	0.330	0.114	52.1	58.6
CHRF1/5	0.332	0.314	58.1	65.0
CHRF1/4	0.334	0.538	63.5	69.5
CHRF1/3	0.338	1.043	69.0	74.3
CHRF1/2	0.347	1.971	75.5	81.9
CHRF1	0.365	3.871	86.2	92.6
CHRF2	**0.370**	**4.400**	**86.7**	**93.6**
CHRF3	0.369	4.286	83.1	91.4
CHRF4	0.368	4.162	80.5	88.6
CHRF5	0.367	4.090	77.6	87.1
CHRF6	0.367	4.081	76.9	87.1
WORDF1/6	0.296	-3.443	6.2	16.6
WORDF1/5	0.296	-3.357	6.9	19.8
WORDF1/4	0.296	-3.348	9.5	21.9
WORDF1/3	0.298	-3.200	16.0	26.9
WORDF1/2	0.300	-2.924	21.9	30.7
WORDF1	0.306	-2.309	31.9	39.8
WORDF2	<u>0.309</u>	<u>-1.995</u>	<u>38.3</u>	<u>47.6</u>
WORDF3	0.308	-2.038	30.2	44.5
WORDF4	0.308	-2.076	28.1	43.1
WORDF5	0.308	-2.090	23.3	39.5
WORDF6	0.308	-2.090	23.8	40.0

Table 1: Overall average segment level (τ) correlation *mean* (column 1), *diff* (column 2), *rank*$>$ (column 3) and *rank*$\geqslant$ (column 4) for each CHRFβ score. Bold represents the overall best value and underline represents the best WORDFβ value. The most promising metric is CHRF2.

- β = 2 is the best option both for CHRF (bold) as well as for WORDF (underline).

Additional observations from Tables 2 and 3:

- for translation into English:
  - the most promising metrics are CHRF2 and CHRF1;
  - the best WORDFβ variant is WORDF2.
- for translation out of English:
  - the most promising metrics are CHRF2 and CHRF3
  - the best WORDFβ variants are WORDF2 and WORDF3

indicating that the recall is even more important for morphologically rich(er) languages.

Regardless to these slight differences between English and non-English texts, CHRF2 can be considered as the most promising variant generally.

metric	*mean*	*diff*	*rank>*	*rank⩾*
CHRF1/6	0.357	1.514	57.6	63.8
CHRF1/5	0.358	1.638	64.8	71.0
CHRF1/4	0.359	1.781	69.0	74.3
CHRF1/3	0.363	2.138	74.8	79.5
CHRF1/2	0.368	2.695	81.9	87.6
CHRF1	0.377	3.695	**91.0**	**98.1**
CHRF2	**0.378**	**3.710**	85.7	91.9
CHRF3	0.376	3.476	83.3	90.0
CHRF4	0.374	3.281	77.6	84.8
CHRF5	0.372	3.091	70.5	78.6
CHRF6	0.372	3.048	70.0	78.1
WORDF1/6	0.308	-3.605	6.2	13.8
WORDF1/5	0.309	-3.481	5.7	16.7
WORDF1/4	0.309	-3.538	10.5	20.5
WORDF1/3	0.311	-3.333	15.2	24.3
WORDF1/2	0.313	-3.076	18.1	24.8
WORDF1	0.320	-2.324	33.3	40.0
WORDF2	0.323	-2.010	40.0	49.0
WORDF3	0.322	-2.143	28.1	41.9
WORDF4	0.322	-2.157	28.1	41.4
WORDF5	0.322	-2.195	23.3	39.0
WORDF6	0.321	-2.205	22.9	38.1

Table 2: Translation into English: average segment level (τ) correlation *mean* (column 1), *diff* (column 2), *rank>* (column 3) and *rank⩾* (column 4) for each CHRFβ score. Bold represents the overall best value and underline represents the best WORDFβ value. The most promising metric is CHRF2.

metric	*mean*	*diff*	*rank>*	*rank⩾*
CHRF1/6	0.290	-1.381	46.0	52.9
CHRF1/5	0.292	-1.138	50.8	58.7
CHRF1/4	0.295	-0.767	57.7	64.6
CHRF1/3	0.302	- 0.138	63.0	68.8
CHRF1/2	0.314	1.186	68.8	74.1
CHRF1	0.342	4.067	82.0	87.8
CHRF2	**0.353**	**5.224**	**87.8**	**94.7**
CHRF3	**0.353**	**5.224**	83.1	93.7
CHRF4	0.352	5.148	83.1	91.5
CHRF5	**0.353**	5.219	83.1	93.7
CHRF6	**0.353**	**5.224**	83.6	93.8
WORDF1/6	0.271	-3.367	6.3	20.1
WORDF1/5	0.271	-3.281	8.4	23.8
WORDF1/4	0.272	-3.267	9.5	25.4
WORDF1/3	0.273	-3.152	16.9	28.6
WORDF1/2	0.276	-2.838	24.3	34.9
WORDF1	0.281	-2.319	29.6	38.1
WORDF2	0.284	-1.976	36.5	45.5
WORDF3	0.285	-1.900	34.4	48.2
WORDF4	0.285	-1.919	29.6	45.5
WORDF5	0.284	-1.929	24.3	40.2
WORDF6	0.285	-1.919	25.9	42.3

Table 3: Translation from English: average segment level (τ) correlation *mean* (column 1), *diff* (column 2), *rank>* (column 3) and *rank⩾* (column 4) for each CHRFβ score. Bold represents the overall best value and underline represents the best WORDFβ value. The most promising metric is CHRF2.

However, taking these differences into account together with the fact that for English, CHRF1 performed better than CHRF3 in the WMT15 metrics shared task, we decided to submit CHRF2 together with CHRF1 and CHRF3 in order to be able to draw more reliable conclusions.

3.1 Investigating n-gram weights for CHRF2

As already mentioned, all CHRFβ variants explored so far are based on the uniform distribution of n-gram weights. Nevertheless, one can assume that character n-grams of different lengths are not equally important – for example, it is conceivable that character 1-grams are not really important for assessment of translation quality. Therefore we carried out the following experiment on the best CHRF variant, namely CHRF2. First step was to examine τ coefficients independently for each n-gram. The results presented in Table **??** indicate that the character 1-grams indeed have the lowest correlation whereas 2-grams and 3-grams have the highest.

Taking these indications into account, we investigated the following three combinations of n-gram weights:

- 0-1-1-1-1-1
  removing 1-grams and keeping uniform weights for the rest of n-grams;
- 1-2-2-2-2-2
  assigning doubled 1-gram weight to the rest of n-grams;
- 1-5-5-4-3-3
  distribution of n-gram weights according to individual n-gram correlations.

The τ coefficients for each n-gram weight distribution are shown in Table 4 – although some of

(a) individual n-grams

n-gram	τ
1-gram	0.280
2-gram	0.361
3-gram	0.367
4-gram	0.358
5-gram	0.347
6-gram	0.334

(b) different n-gram weight distributions

Kendall's τ	fr-en		de-en		cs-en		ru-en		hi-en	fi-en	avg.
011111	**.397**	.384	.320	.424	.266	.437	.317	.385	.396	.406	.373
122222	.395	**.385**	.325	**.425**	.270	.451	.318	.389	.405	.405	.377
155433	.396	**.385**	.327	**.425**	.274	.451	.319	.388	.403	**.407**	.377
uniform	.394	.381	**.331**	.424	**.275.**	**.451**	**.320**	**.394**	**.410**	.398	***.378***

Kendall's τ	en-fr		en-de		en-cs		en-ru		en-hi	en-fi	avg.
011111	.300	**.345**	.256	.382	.334	.441	**.460**	**.420**	.304	.359	.360
122222	.302	.338	.261	.388	**.336**	.445	.457	.418	.304	.366	.361
155433	**.303**	.342	.260	.387	**.336**	**.449**	.456	.419	.305	.366	.362
uniform	.302	.338	**.264**	**.393**	.334	.444	.453	.418	**.307**	**.375**	***.363***

Table 4: Analysis of n-grams: (a) average τ for individual n-grams (b) τ on WMT14 (left) and WMT15 (right) documents for different n-gram weight distributions.

the proposed distributions outperform the uniform one for some of the texts, especially for translation out of English, none of them is unquestionably better than the uniform distribution of weights.

Therefore, the uniform n-gram weights were used for the WMT16 metrics task.

4 CHRF and WORDF for good and bad translations

In order to try to better understand the differences between WORDF and CHRF scores, i.e. the advantages of the CHRF score, we carried out a preliminary experiment on three data sets for which the absolute (direct) human scores were available. The data sets are rather heterogeneous: they contain three different target languages, they were produced and evaluated independently, for different purposes, and the human scores were not defined in the same way. In addition, two of the three data sets are rather small. Therefore the described experiment is rather preliminary, however we believe that it represents a good starting point for further research regarding differences between word and character based metrics.

τ coefficients for comparing four systems using direct human scores

The starting point was testing τ coefficients for CHRF2 and WORDF2 on the English→Spanish data set described in (Specia et al., 2010) and the motivation was simply to explore the correlations obtained on direct human scores instead of relative rankings. The data set contains 4000 source segments and their reference translations, machine translation outputs of four SMT systems, as well as human estimations of required post-editing effort in the interval from 1 (requires complete retranslation) to 4 (fit for purpose). The distribution of segments with each of the four human ratings for each of the systems is shown in Table 5a and it can be seen that the fourth system is significantly worse than the other three, which are rather close.

The obtained τ coefficients (Table 5b, first column) were however puzzling – the τ coefficients are very close, the one for the WORDF2 is even slightly higher, which is a rather different result than all the results described in the previous sections and related work. On the other hand, taking into account that the number of systems is small, as well as that the performance of the fourth sys-

(a) Distribution of direct human scores

human score	1	2	3	4	mean
sys1	4.2	24.8	54.3	16.7	2.83
sys2	8.9	36.5	44.4	10.2	2.56
sys3	9.7	38.5	43.2	8.6	2.51
sys4	73.0	20.6	5.9	0.5	1.34

(b) τ correlations

τ	4 sys	3 sys
WORDF2	**0.615**	0.275
CHRF2	0.608	**0.313**

Table 5: English→Spanish data set with direct human scores: (a) percentage of the sentence level human scores for each of the four systems together with the average human score for each system – system 4 is significantly worse than the other three. (b) τ coefficients for all four systems (first column) and for the three similar systems (second column).

tem is clearly distinct than of the others, another experiment is carried out: the worst system is removed and only the remaining three similar systems are compared. For this set-up, the expected results were obtained (second column), i.e. the τ coefficients are higher for the CHRF2 score. This somewhat controversial finding lead to the following two hypotheses:

1. word-based metrics are good at distinguishing systems/segments of distinct quality but not so good at ranking similar systems/segments;

2. word-based metrics are good for evaluating low quality systems/segments but not so good for evaluating high quality systems/segments.

Standard deviations of automatic metrics for different direct human scores

In order to further examine the two hypotheses, the following experiment has been carried out: for each of the human ratings, standard deviation of the corresponding automatic scores is calculated. This experiment is carried out on the previously described data set as well as on two additional small[1] data sets:

- English→Irish SMT translations rated from 1 to 4 for the overall quality (1=bad, 4=good);

[1] about 200 segments

(a) English→Spanish

hum	WORDF2	CHRF2
1	10.4	11.5
2	12.8	12.1
3	15.8	14.2
4	21.7	17.3

(b) English→Irish

hum	WORDF2	CHRF2
1	7.7	7.0
2	8.1	9.6
3	6.3	4.3
4	24.3	14.0

(c) English→Serbian

hum	WORDF2	CHRF2
1	6.8	8.9
1.5	4.6	6.4
2	11.2	9.9
2.5	13.4	11.1
3	13.2	11.5
3.5	11.2	8.3
4	15.4	9.9
4.5	16.4	7.4
5	25.0	7.7

Table 6: Standard deviations of WORDF2 and CHRF2 for each value of direct human scores on three distinct datasets: (a) English→Spanish, estimated post-editing effort (b) English→Irish, overall quality (c) English→Serbian, average of adequacy and fluency.

- English→Serbian SMT translations rated from 1 to 5 in terms of adequacy and fluency (1=bad, 5=good) – the mean value of the two has been taken as the direct human score.

The obtained standard deviations in Table 6 show that for poorly rated sentences, the deviations of CHRF2 and WORDF2 are similar – both metrics assign relatively similar (low) scores. On the other hand, for the sentences with higher human rates, the deviations for CHRF2 are (much) lower. In addition, the higher the human rating is, the greater is the difference between the WORDF2 and CHRF2 deviations. These results confirm the hypothesis 2), namely that CHRF is better than WORDF mainly for segments/systems of higher translation quality. The most probable reason is that CHRF, contrary to the word-based metrics, does not penalise too hard acceptable morpho-

syntactic variations. The CHRF scores for good translations are therefore more concentrated in the higher range, whereas the WORDF scores are often too low. The results are also consistent with the hypothesis 1), however this one is confirmed only partially since the outlier is a low quality system – further work should include comparison of different low quality systems.

Nevertheless, as stated at the beginning of the section, it should be kept in mind that this is only a preliminary experiment in this direction, performed on very limited amount of data. Further experiments on large data sets, more systems and more languages should be carried out in order to get more reliable results and better insight into underlying phenomena.

5 Summary and outlook

The results presented in this work show that generally, the F-scores which are biased towards recall correlate better with human rankings than those biased towards precision. Particularly, it is shown that CHRF2 version of the CHRF score with uniform n-gram weights is the most promising for machine translation evaluation. Therefore this/these version has been submitted to the WMT16 metrics task, however together with CHRF1 and CHRF3 in order to explore differences between English and morphologically richer target languages more systematically.

In addition, it is shown that the CHRF score performs better than the WORDF score. Preliminary experiments on small data sets with available direct human scores show that for sentences of higher translation quality, standard deviations of WORDF is much larger than standard deviations of CHRF, indicating that the main advantage of the CHRF is that it does not penalise too strong different variants of acceptable translations. However, more systematic experiments on large data sets should be carried out in this direction. Furthermore, a broader investigation including different word and character based metric in addition to the two presented F-scores would be useful.

Apart from this, application of CHRF on more distinct languages such as Arabic, Chinese etc. should be explored.

Acknowledgments

This work emerged from research supported by 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 Mihael Arčan and Sanja Štajner for providing additional small data sets with direct human scores.

References

Ondrej 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 9th Workshop on Statistical Machine Translation (WMT-14)*, pages 12–58, Baltimore, Maryland, USA, June.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*, pages 1–46, Lisbon, Portugal, September.

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

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)*, pages 392–395, Lisbon, Portugal, September.

Lucia Specia, Nicola Cancedda, and Marc Dymetman. 2010. A dataset for assessing machine translation evaluation metrics. In *Seventh Conference on International Language Resources and Evaluation*, LREC, pages 3375–3378, Valletta, Malta.

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)*, pages 256–273, Lisbon, Portugal, September.