PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 98 OCTOBER 2012 99–108

rgbF: An Open Source Tool for n-gram Based Automatic Evaluation of Machine Translation Output

Maja Popović

DFKI, Language Technology Group

Abstract

We describe rgbF, a tool for automatic evaluation of machine translation output based on n-gram precision and recall. The tool calculates the F-score averaged on all n-grams of an arbitrary set of distinct units such as words, morphemes, POS tags, etc. The arithmetic mean is used for n-gram averaging. As input, the tool requires reference translation(s) and hypothesis, both containing the same combination of units. The default output is the document level 4-gram F-score of the desired unit combination. The scores at the sentence level can be obtained on demand, as well as precision and/or recall scores, separate unit scores and separate n-gram scores. In addition, weights can be introduced both for n-grams and for units, as well as the desired n-gram order n.

1. Motivation

Evaluation of machine translation output is an important but difficult task. A number of automatic evaluation measures have been studied over the years, some of them have become widely used by machine translation researchers, such as the Bleu metric (Papineni et al., 2002) and the Translation Edit Distance TER (Snover et al., 2006). Precision and recall are used for machine translation evaluation in Melamed et al. (2003) and it is shown that they correlate well with human judgments, even better than the BLEU score. Recent investigations have shown that the n-gram based evaluation metrics BLEU and F-score calculated on Part-of-Speech (POS) sequences correlate very well with human judgments (Popović and Ney, 2009) clearly outperforming the widely used metrics BLEU and TER. However, using only POS tags for evaluation has

Corresponding author: maja.popovic@dfki.de
Cite as: Maja Popović. rgbF: An Open Source Tool for n-gram Based Automatic Evaluation of Machine Translation Output. The Prague Bulletin of Mathematical Linguistics No. 98, 2012, pp. 99–108.
doi: 10.2478/v10108-012-0012-y.

certain disadvantages, for example the translation hypotheses "The flowers are beautiful" and "The results are good" would have the same score. Therefore combining lexical and non-lexical "units", e.g. words and POS tags seemed to be a promising direction for further investigation.

The RGBF tool presented in this work enables calculation of such combined scores, i.e. F-score of an arbitrary combination of distinct units (words, POS tags, morphemes, etc). The tool has been successfully used in the sixth WMT evaluation shared task (Popović, 2011; Callison-Burch et al., 2011), and it is confirmed that introducing the morphological and syntactic properties of involved languages thus abstracting away from word surface particularities (such as vocabulary and domain) improves the correlation with human judgments, especially for the translation from English.

The name RGBF refers to the RGB color model used in computer graphics: in this model, primary colors red, green, and blue are added together in various ways thus producing a broad array of different colors. Our evaluation tool adds together individual scores for different basic units and n-gram orders in various ways thus producing a broad array of evaluation scores. The final letter F stands for the F-score which is the default output.

The tool is written in Python, and it is available under an open-source licence. We hope that the release of the toolkit will facilitate the automatic evaluation for the researchers, and also stimulate further development of the proposed method.

2. RGBF tool

2.1. Algorithm

RGBF implements the precision, recall and F-score of all n-grams up to order n of all desired units. The arithmetic averaging of n-grams is performed – previous experiments on the syntax-oriented n-gram metrics (Popović and Ney, 2009) showed that there is no significant difference between arithmetic and geometric mean in the terms of correlation coefficients. In addition, it is also argued that the geometric mean used for the BLEU score is not optimal because the score becomes equal to zero even if only one of the n-gram counts is equal to zero, which is especially problematic for the sentence level evaluation.

The recall is defined as percentage of words in the reference which also appear in the hypothesis, and analogously, the precision is the percentage of words in the hypothesis which also appear in the reference. Multiple counting is not allowed. For example, for the hypothesis "this is a hypothesis and this is a hypothesis" and the reference "this is a reference and this is a hypothesis" the unigram precision will be 8/9=88.9% and not 9/9=100%. In the case of multiple references, the highest precision and the highest recall score is chosen for each sentence (the optimal reference for the precision and the optimal reference for the recall are not necessarily the same). Once the recall and precision are obtained, the F-score is calculated as their harmonic mean.

Although the method is generally language-independent, availability of some kind of analyser for the particular target language might be required depending on which units are desired.

2.2. Usage

RGBF supports the option `-h/--help` which outputs a description of the available command line options.

The input options are:

`-R, --ref`	translation reference
`-H, --hyp`	translation hypothesis
`-n, --ngram`	n-gram order (default: $n = 4$)
`-uw, --uweight`	unit weights (default: $1/U$)
`-nw, --nweight`	n-gram weights (default: $1/n$)

Inputs `-R` and `-H` are required, containing an arbitrary number of different types of units. The combination of units must be the same and in the same order both in the reference and in the hypothesis, and the units must be separated by "++". This symbol is of course not needed if the input files contain only one unit. The required format for all input files is a raw tokenized text containing one sentence per line. In the case of multiple references, all available reference sentences must be separated by the symbol #.

The output options are:

- standard output – the default output of the tool is the overall (document level) 4-gram F-score.

In addition to the standard output, the following optional outputs are available:

`-p, --prec`	precision
`-r, --rec`	recall
`-u, --unit`	separate unit scores
`-g, --gram`	separate n-gram scores
`-s, --sent`	sentence level scores

An example of input and output files and different program calls is shown in the next section.

2.3. Example

Table 1 presents an example of translation hypothesis consisting of two sentences and its corresponding reference translation in the RGBF format. Both hypothesis and reference contain four types of units, i.e. full words, base forms, morphemes and pos tags, separated by ″++″.

`example.hyp.wbmp (word+base+morph+pos)`
This time , the reason for the collapse on Wall Street . ++ This time , the reason for the collapse on Wall Street . ++ Th is time , the reason for the collapse on Wall Street . ++ DT NN , DT NN IN DT NN IN NP NP SENT
The proper functioning of the market and a price . ++ The proper functioning of the market and a price . ++ The proper function ing of the market and a price . ++ DT JJ NN IN DT NN CC DT NN SENT

`example.ref.wbmp (word+base+morph+pos)`
This time the fall in stocks on Wall Street is responsible for the drop . ++ This time the fall in stock on Wall Street be responsible for the drop . ++ Th is time the fall in stock s on Wall Street is responsible for the drop . ++ DT NN DT NN IN NNS IN NP NP VBZ JJ IN DT NN SENT
The proper functioning of the market environment and the decrease in prices . ++ The proper functioning of the market environment and the decrease in price . ++ The proper function ing of the market environment and the decrease in price s . ++ DT JJ NN IN DT NN NN CC DT NN IN NNS SENT

Table 1. Example of a hypothesis and a corresponding reference containing four units: full words, base forms, morphemes and pos *tags merged in the* rgbF *format.*

1) *Simple program call* without optional parameters:

```
rgbF.py -R example.ref.wbmp -H example.hyp.wbmp
```

will calculate the document level F-score with the default n-gram order $n = 4$ and the uniform distribution of weights, i.e. all the n-gram weights are $1/n = 1/4 = 0.25$ and all the unit weights are $1/U$ where U is the number of different units ($U = 4$ for the input files presented in Table 1). The obtained output will be:

```
rgbF    42.2512
```

2) A *desired unit and/or n-gram weight distribution* can be demanded with a call:

```
rgbF.py -R example.ref.wbmp -H example.hyp.wbmp -uw 2-3-4-6 -nw 2-2-5-5
```

where `uw` represents the proportion of unit weights and `nw` the proportion of n-gram weights. The weights are normalized automatically, so that the given numbers do not have to sum to 1, only to represent the desired proportion. The output of this call will be:

```
rgbF    36.5530
```

3) The weights also enable *the choice of units and/or n-grams*. For example, the call:

```
rgbF.py -R example.ref.wbmp -H example.hyp.wbmp -uw 2-0-0-3
```

will produce the word+POS F-score averaged on unigrams, bigrams, trigrams and fourgrams in proportion 2 words : 3 POS, and the call:

```
rgbF.py -R example.ref.wbmp -H example.hyp.wbmp -nw 1-0-0-1
```

will average over all units but only over unigrams and fourgrams.

4) A *desired maximum n-gram order* can also be demanded, for example 6-gram:

```
rgbF.py -R example.ref.wbmp -H example.hyp.wbmp -n 6
```

5) *Precision and/or recall scores* can be requested:

```
rgbF.py -R example.ref.wbmp -H example.hyp.wbmp -p -r
```

These scores will be then showed in addition to the default F-score:

```
rgbF      42.2512
rgbPrec   48.9473
rgbRec    37.1839
```

6) If *the sentence scores* are desired:

```
rgbF.py -R example.ref.wbmp -H example.hyp.wbmp -s
```

the F-score of each sentence together with the sentence number will be showed in addition to the default document level F-score:

```
1::rgbF   31.0037
2::rgbF   55.8205
rgbF      42.2512
```

7) If *the unit scores* are demanded:

```
rgbF.py -R example.ref.wbmp -H example.hyp.wbmp -u
```

the F-score of each unit will be showed in addition to the default overall F-score:

```
u1-F   36.6824
u2-F   38.7693
u3-F   40.2712
u4-F   53.2818
rgbF   42.2512
```

where the unit number is its position in the reference and hypothesis file. For our example, u1 stands for the full words, u2 for base forms, u3 are morphemes and u4 are POS tags.

8) Separate n-*gram scores* can also be demanded:

```
rgbF.py -R example.ref.wbmp -H example.hyp.wbmp -g
```

so that the F-score of each n-gram of each unit will be showed in addition to the default overall F-score:

```
u1-1gram-F   68.0000
u1-2gram-F   39.1304
u1-3gram-F   23.8095
u1-4gram-F   15.7895
u2-1gram-F   72.0000
u2-2gram-F   43.4783
...          ...
u4-3gram-F   42.8571
u4-4gram-F   21.0526
rgbF         42.2512
```

9) The *most "complicated" program call* involving *all optional output parameters*:

```
rgbF.py -R example.ref.wbmp -H example.hyp.wbmp -p -r -u -g -s
```

will produce all the F-scores, precisions and recalls for each unit n-gram and each unit, on the sentence level and on the document level.

3. Correlations with human ranking

As mentioned in Section 1, the tool has been tested on all WMT data from year 2008 to year 2011. In addition, it has also been tested on the data developed in the framework of the TARAXÜ project[1]. Spearman's rank correlation coefficients ρ are calculated for the document (system) level correlation, whereas Kendall's τ coefficients are calculated for the sentence level correlation.

3.1. WMT data

The following 4-gram RGBF scores have been investigated on the WMT data: WORDF, MORPHF, POSF, WPF, WMF, MPF, as well as WMPF without and with given weights (WMPF'). Spearman's rank correlation coefficients on the document (system) level between all the metrics and the human ranking are computed on the English, French, Spanish, German and Czech texts generated by various translation systems in the framework of the third, fourth and fifth shared translation tasks (Callison-Burch et al., 2008, 2009, 2010), and the results are shown in Table 2.

[1] http://taraxu.dfki.de/

metric	overall	x→en	en→x
BLEU	0.566	0.587	0.544
WORDF	0.550	0.592	0.504
MORPHF	0.608	0.671	0.541
POSF	**0.673**	**0.726**	**0.617**
WPF	0.627	0.698	0.553
WMF	0.587	0.655	0.514
MPF	**0.669**	**0.744**	**0.590**
WMPF	0.645	0.721	0.565
WMPF′	**0.668**	**0.744**	**0.587**

Table 2. Average document level correlations on the WMT *2008–2010 data for the* BLEU *score and the investigated* RGB *metrics. Bold represents the best value in the particular metric group (single unit, two-unit and three-unit). The most promising metrics are those containing* POS *and morpheme information, namely* WMPF′ *(*WMPF *with non-uniform weights),* MPF *and* POSF.

The most promising metrics, i.e. MPF and WMPF′ are submitted to the sixth shared evaluation task (Callison-Burch et al., 2011) and the correlations on the document and on the sentence level are presented in Table 3, together with the widely used BLEU and TER metrics and the best ranked metrics MTERATERPLUS, TINE-SRL-MATCH, TESLA-F, TESLA-B, METEOR-adq, METEOR-rank and AMBER.

On the document level, the RGBF scores are better than BLEU and TER and comparable with the best ranked metrics for translation from English, however worse than the best ranked metrics for translation into English. On the sentence level, the RGBF scores are comparable with the best ranked metrics for translation into English, and one of the best for translation from English.

3.2. TARAXÜ data

The TARAXÜ corpora consist of two domains: News taken from the WMT 2010 News test set and technical documentation extracted from the freely available OpenOffice project (Tiedemann, 2009). The translation outputs are produced by four different German-to-English, English-to-German and Spanish-to-German machine translation systems: Google, Moses (statistical systems), Lucy (a rule-based system) and Trados (not really a system but a translation memory). The obtained outputs are then given to the professional human annotators to assign 1–4 ranks, but without ties. More details can be found in (Avramidis et al., 2012).

The following 4-gram RGB scores have been explored on this data: WORDF, BASEF, MORPHF, POSF, WPF, BPF, MPF, WMPF, MBPF and WMBPF, all with the default uniform weights.

	document level		sentence level	
metric	x→en	en→x	x→en	en→x
MPF	0.77	0.78	0.28	0.26
WMPF	0.76	0.77	0.27	0.25
BLEU	0.69	0.70	/	/
TER	0.67	0.57	/	/
MTeRaterPlus	0.90	/	0.37	/
TINE-SRL-MATCH	0.87	/	0.23	/
TESLA-F	0.86	0.94*	0.31	0.26*
TESLA-B	0.84	0.87*	0.30	0.25*
METEOR-adq	0.83	/	0.28	/
METEOR-rank	0.82	0.63	0.29	0.23
AMBER	0.80	0.70	0.27	0.26

*Table 3. Average document level and sentence level correlations on WMT 2011 shared evaluation task for two submitted RGB metrics, widely used BLEU and TER scores, and best ranked novel evaluation metrics. The results marked with * are averaged without the Czech translation outputs.*

Document level Spearman's coefficients and sentence level Kendall's coefficients are calculated for the BLEU score and for all investigated RGBF scores on all data, as well as separately for each language pair and for each domain.

On the document level no significant differences are observed – all the correlation coefficients are very high, between 0.8 and 1. Sentence level correlations are shown in Table 4. The results are similar to those on WMT data, i.e. most promising metric is the MPF score, followed by the WMPF and MBPF scores. Combining full forms and base forms of the words (WMBPF) does not yield any improvements.

4. Conclusions

We presented RGBF, a toolkit for automatic evaluation of translation output which we believe will be of value to the machine translation community. It can be downloaded from `http://www.dfki.de/~mapo02/rgbF/`.

So far, the most promising RGBF scores are those using morphemes and POS tags as units. Different unit and n-gram weights should be investigated in future work, as well as the use of other types of units.

	overall	de-en	en-de	es-de	news	openoffice
bleu	-0.198	0.024	-0.250	-0.296	-0.181	-0.328
wordF	0.557	0.592	0.544	0.544	0.549	0.619
baseF	0.561	0.589	0.554	0.548	0.553	0.618
morphF	0.587	0.616	0.570	0.583	0.581	0.639
posF	0.534	0.569	0.511	0.529	0.528	0.582
wpF	0.577	0.610	0.564	0.565	0.571	0.624
bpF	0.577	0.611	0.563	0.566	0.571	0.622
mpF	**0.597**	**0.623**	0.587	**0.589**	**0.591**	0.644
wmpF	0.595	0.622	0.582	0.587	0.588	0.645
mbpF	0.596	0.620	**0.589**	0.588	0.589	**0.654**
wmbpF	0.593	0.618	0.583	0.586	0.586	0.650

Table 4. Sentence level correlations on taraXÜ data for the bleu score and the investigated rgb metrics. Bold represents the best values. The most promising metrics are mpF, wmpF and mbpF.

Acknowledgments

This work has partly been developed within the taraXÜ project financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development.

Bibliography

Avramidis, Eleftherios, Aljoscha Burchardt, Christian Federmann, Maja Popović, Cindy Tscherwinka, and David Vilar. Involving language professionals in the evaluation of machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC12)*, Istanbul, Turkey, May 2012.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT 2008)*, pages 70–106, Columbus, Ohio, June 2008.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT 2009)*, pages 1–28, Athens, Greece, March 2009.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR (WMT 2010)*, pages 17–53, Uppsala, Sweden, July 2010.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 22–64, Edinburgh, Scotland, July 2011.

Melamed, I. Dan, Ryan Green, and Joseph P. Turian. Precision and Recall of Machine Translation. In *Proceedings of the Human Language Technology Conference (HLT-NAACL 03)*, pages 61–63, Edmonton, Canada, May/June 2003.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wie-Jing Zhu. 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 2002.

Popović, Maja. 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 2011.

Popović, Maja and Hermann Ney. Syntax-oriented evaluation measures for machine translation output. In *Proceedings of the 4th EACL 09 Workshop on Statistical Machine Translation (WMT 2009)*, pages 29–32, Athens, Greece, March 2009.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 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 2006.

Tiedemann, Jorg. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins Amsterdam, Borovets, Bulgaria, 2009.

Address for correspondence:
Maja Popović
maja.popovic@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Language Technology Group (LT)
Alt-Moabit 91c
10559 Berlin, Germany