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
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

rbgF 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 rbgF format. Both hypothesis and refer-
ence 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 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 functioning 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 functioning 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+ros F-score averaged on unigrams, bigrams, trigrams and fourgrams in proportion 2 words : 3 ros, 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 ros tags.
8) Separate \textit{n-gram scores} can also be demanded:

\texttt{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:

\begin{verbatim}
  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
\end{verbatim}

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

\texttt{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 \textit{wmt} data from year 2008 to year 2011. In addition, it has also been tested on the data developed in the framework of the \textit{taraxu} project\textsuperscript{1}. Spearman's rank correlation coefficients $\rho$ are calculated for the document (system) level correlation, whereas Kendall's $\tau$ coefficients are calculated for the sentence level correlation.

3.1. \textit{wmt} data

The following 4-gram \texttt{rgbF} scores have been investigated on the \textit{wmt} data: \texttt{wordF}, \texttt{morphF}, \texttt{posF}, \texttt{wpF}, \texttt{wmF}, \texttt{mpF}, as well as \texttt{wmpF} without and with given weights (\texttt{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 (\textit{Callison-Burch et al., 2008, 2009, 2010}), and the results are shown in Table 2.

\textsuperscript{1}http://taraxu.dfk.de/
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.

<table>
<thead>
<tr>
<th>metric</th>
<th>overall</th>
<th>x→en</th>
<th>en→x</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>0.566</td>
<td>0.587</td>
<td>0.544</td>
</tr>
<tr>
<td>WORDF</td>
<td>0.550</td>
<td>0.592</td>
<td>0.504</td>
</tr>
<tr>
<td>MORPHF</td>
<td>0.608</td>
<td>0.671</td>
<td>0.541</td>
</tr>
<tr>
<td>POSF</td>
<td>0.673</td>
<td>0.726</td>
<td>0.617</td>
</tr>
<tr>
<td>WPF</td>
<td>0.627</td>
<td>0.698</td>
<td>0.553</td>
</tr>
<tr>
<td>WMF</td>
<td>0.587</td>
<td>0.655</td>
<td>0.514</td>
</tr>
<tr>
<td>MPF</td>
<td>0.669</td>
<td>0.744</td>
<td>0.590</td>
</tr>
<tr>
<td>WMPF</td>
<td>0.645</td>
<td>0.721</td>
<td>0.565</td>
</tr>
<tr>
<td>WMPF’</td>
<td>0.668</td>
<td>0.744</td>
<td>0.587</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>metric</th>
<th>document level</th>
<th>sentence level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x→en</td>
<td>en→x</td>
</tr>
<tr>
<td>mpF</td>
<td>0.77</td>
<td>0.78</td>
</tr>
<tr>
<td>wmpF</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>BLEU</td>
<td>0.69</td>
<td>0.70</td>
</tr>
<tr>
<td>TER</td>
<td>0.67</td>
<td>0.57</td>
</tr>
<tr>
<td>MTerraterPlus</td>
<td>0.90</td>
<td>/</td>
</tr>
<tr>
<td>TINE-srl-match</td>
<td>0.87</td>
<td>/</td>
</tr>
<tr>
<td>TESLA-F</td>
<td>0.86</td>
<td>0.94*</td>
</tr>
<tr>
<td>TESLA-B</td>
<td>0.84</td>
<td>0.87*</td>
</tr>
<tr>
<td>METEOR-adq</td>
<td>0.83</td>
<td>/</td>
</tr>
<tr>
<td>METEOR-rank</td>
<td>0.82</td>
<td>0.63</td>
</tr>
<tr>
<td>AMBER</td>
<td>0.80</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Document level Spearman’s coefficients and sentence level Kendall’s coefficients are calculated for the BLEU score and for all investigated RCB 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 (mbpF) does not yield any improvements.

4. Conclusions

We presented RCB, 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 RCB 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.
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


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