Morphemes and POS tags for n-gram based evaluation metrics

Maja Popović
German Research Center for Artificial Intelligence (DFKI)
Language Technology (LT), Berlin, Germany
maja.popovic@dfki.de

Abstract

We propose the use of morphemes for automatic evaluation of machine translation output, and systematically investigate a set of F score and BLEU score based metrics calculated on words, morphemes and POS tags along with all corresponding combinations. Correlations between the new metrics and human judgments are calculated on the data of the third, fourth and fifth shared tasks of the Statistical Machine Translation Workshop. Machine translation outputs in five different European languages are used: English, Spanish, French, German and Czech. The results show that the F scores which take into account morphemes and POS tags are the most promising metrics.

1 Introduction

Recent investigations have shown that the n-gram based evaluation metrics calculated on Part-of-Speech (POS) sequences correlate very well with human judgments (Callison-Burch et al., 2008; Callison-Burch et al., 2009; Popović and Ney, 2009) clearly outperforming the widely used metrics BLEU and TER. The BLEU score measured on morphemes is shown to be useful for evaluation of morphologically rich languages (Luong et al., 2010). We propose the use of morphemes for a set of n-gram based automatic evaluation metrics and investigate the correlation of the novel metrics with human judgments. We carry out a systematic comparison between the F and BLEU based metrics calculated on various combinations of words, morphemes and POS tags. The focus of this work is not a comparison of the morpheme and POS based metrics with the standard evaluation metrics[1] as in (Popović and Ney, 2009), but rather a comparison within the proposed set of metrics in order to decide which score(s) should be submitted to the WMT 2011 evaluation task. There are fifteen evaluation metrics in total, which can be divided in three groups: the metrics calculated on single units, i.e. words, morphemes or POS tags alone, the metrics calculated on pairs, i.e. words and POS tags, words and morphemes as well as morphemes and POS tags, and the metrics which take everything into account – lexical, morphological and syntactic information, i.e. words, morphemes and POS tags.

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 (Callison-Burch et al., 2008), fourth (Callison-Burch et al., 2009) and fifth (Callison-Burch et al., 2010) shared translation tasks.

2 Evaluation metrics

We carried out a systematic comparison between the following metrics:

- single unit (word/morpheme/POS) metrics:
  - WORDF
    Standard F score: takes into account all word n-grams which have a counterpart

[1] Apart from the standard BLEU score which is tightly related.

both in the corresponding reference and in the hypothesis.

  - MORPHF
    Morpheme F score: takes into account all morpheme n-grams which have a counterpart both in the corresponding reference and in the hypothesis.
  - POSF
    POS F score: takes into account all POS n-grams which have a counterpart both in the corresponding reference and in the hypothesis.
  - BLEU
    The standard BLEU score (Papineni et al., 2002).
  - POSBLEU
    The standard BLEU score calculated on POS tags.
  - MORPHBLEU
    The standard BLEU score calculated on morphemes.

- pairwise metrics:
  - WPF
    F score of word and POS n-grams.
  - WMF
    F score of word and morpheme n-grams.
  - MPF
    F score of morpheme and POS n-grams.
  - WPBLEU
    Arithmetic mean of BLEU and POSBLEU scores.
  - WMBLEU
    Arithmetic mean of BLEU and MORPHBLEU scores.
  - MPBLEU
    Arithmetic mean of MORPHBLEU and POSBLEU scores.

- metrics taking everything into account:
  - WMPF
    F score on word, morpheme and POS n-grams.
  - WMPBLEU
    Arithmetic mean of BLEU, MORPHBLEU and POSBLEU scores.
  - WMPFBLEU
    Arithmetic mean of all F and BLEU scores.

The prerequisite for POS based metrics is availability of an appropriate POS tagger for the target language. It should be noted that the POS tags cannot be only basic but must have all details (e.g. verb tenses, cases, number, gender, etc.). For the morpheme based metrics, a tool for splitting words into morphemes is necessary.

All the F scores and the BLEU scores are based on four-grams (i.e. the value of maximal n is 4). Preliminary experiments on the morpheme based measures showed that there is no improvement by using six-grams, seven-grams or eight-grams. As for the n-gram averaging, BLEU scores use geometric mean. However, it is also argued not to be optimal because the score becomes equal to zero even if only one of the n-gram counts is equal to zero. In addition, 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. Therefore, arithmetic averaging without weights is used for all F-scores. For the WMPF score, an additional experiment with weights is carried out as well.

3 Experiments on WMT 2008, WMT 2009 and WMT 2010 test data

Experimental set-up

The evaluation metrics were compared with human rankings by means of Spearman correlation coefficients ρ. Spearman's rank correlation coefficient is equivalent to Pearson correlation on ranks, and its advantage is that it makes fewer assumptions about the data. The possible values of ρ range between 1 (if all systems are ranked in the same order) and -1 (if all systems are ranked in the reverse order). Thus the higher the value of ρ for an automatic metric, the more similar is to the human metric.

The scores were calculated for outputs of translations from Spanish, French, German and Czech into English and vice versa. Spanish, French, German and English POS tags were produced using the TreeTagger[2], and the Czech texts are tagged using the

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

COMPOST tagger (Spoustová et al., 2009). In this way, all references and hypotheses were provided with detailed POS tags.

The words of all outputs were split into morphemes using the Morfessor tool (Creutz and Lagus, 2005). The tool is corpus-based and language-independent: it takes a text as input and produces a segmentation of the word forms observed in the text. The obtained results are not strictly linguistic, however they often resemble a linguistic morpheme segmentation. Once a morpheme segmentation has been learnt from some text, it can be used for segmenting new texts. In our experiments, for each document, first a corresponding reference translation has been split, and then this segmentation is used for splitting all translation hypotheses. In this way, possible discrepancies between reference and hypothesis segmentation of the same word are avoided. Effects of the training on the large(r) monolingual corpora have not been investigated yet.

In Table 1, an English reference sentence can be seen along with its morpheme and POS equivalents.

words	Another leading role in the film is played by Matt Damon .
morphemes	An other lead ing role in the film is play ed by Ma tt Da mon .
POS tags	DT VBG NN IN DT NN VBZ VBN IN NP NP SENT

Table 1: Example of an English sentence with its corresponding morpheme and POS sequences.

Comparison of metrics

For each evaluation metric described in Section 2, the system level Spearman correlation coefficients ρ were calculated for each document. In total, 33 correlation coefficients were obtained for each metric – four English outputs from the WMT 2010 task, five from the WMT 2009 and eight from the WMT 2008 task, together with sixteen outputs in other four target languages. The obtained correlation results were then summarised into the following three values:

- *mean*
  a correlation coefficient averaged over all translation outputs;

- *rank>*
  percentage of documents where the particular metric has better correlation than the other metrics investigated in this work;

- *rank≥*
  percentage of documents where the particular metric has better or equal correlation than the other metrics investigated in this work.

These values for each metric are presented in Table 2.

metric	*mean*	*rank>*	*rank≥*
WORDF	0.550	24.2	42.6
MORPHF	0.608	40.0	58.0
POSF	**0.673**	**63.4**	**78.0**
BLEU	0.566	20.6	38.6
MORPHBLEU	0.567	29.9	44.6
POSBLEU	**0.674**	**54.7**	**66.9**
WPF	0.627	44.0	66.9
WMF	0.587	37.0	53.9
MPF	**0.669**	**51.9**	**77.4**
WPBLEU	0.629	41.0	57.4
WMBLEU	0.557	23.6	41.0
MPBLEU	**0.634**	**44.6**	**66.6**
WMPF	0.645	46.3	71.1
WMPBLEU	0.610	32.7	54.7
WMPFBLEU	0.628	35.8	61.6
WMPF’	**0.668**	**51.9**	**78.8**

Table 2: Average correlation *mean* (column 1), *rank>* (column 2) and *rank≥* (column 3) for each evaluation metric. Bold represents the best value in the particular metric group. The most promising metrics are the F scores containing POS and morpheme information, namely WMPF’, MPF and POSF, as well as the POSBLEU score. The standard BLEU score has very low values.

It can be observed that the morpheme based metrics outperform the word based metrics, however not the POS based metrics. As for pairwise metrics, the MPF score seems to be very promising. Adding the actual original words unfortunately deteriorates the system level correlations, nevertheless omitting the words can possibly lead to the poor sentence level correlations. Therefore an additional experiment is carried out with the most promising metric containing words, namely the WMPF score: a weighted

WMPF' score is introduced, with word weight of 0.2, morpheme weight of 0.3 and POS weight of 0.5. WMPF' clearly outperforms the simple WMPF score without weights, and it is comparable to the morpheme-POS F score MPF as well as POS-based metrics POSF and POSBLEU. Apart from that, it can be observed that, in general, the F scores are better than the BLEU scores. The combination of all F and all BLEU scores (WMPFBLEU) is better than the WMPBLEU score, but does not yield any improvements over the WMPF score.

The most promising metrics are the F scores containing POS and morpheme information, namely POSF, MPF and WMPF' together with the WMPF, as well as the POSBLEU score. The standard BLEU score has the third lowest average correlation and the lowest rank values.

4 Conclusions

The results presented in this article show that the use of morphemes improves n-gram based automatic evaluation metrics, particularly in combination with syntactic information in the form of detailed POS tags. Especially promising are the weighted WMPF and the MPF scores, which have been submitted to the WMT 2011 evaluation task. Weights for these two metrics should be further investigated in future work, as well as the possible impact of different morpheme splittings (such as training on larger texts).

Acknowledgments

This work has partly been developed within the TARAXÜ project[3] financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development.

References

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)*, pages 70–106, Columbus, Ohio, June.

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

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. 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 10)*, pages 17–53, Uppsala, Sweden, July.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Technical Report Report A81, Computer and Information Science, Helsinki University of Technology, Helsinki, Finland, March.

Minh-Thang Luong, Preslav Nakov, and Min-Yen Kan. 2010. A Hybrid Morpheme-Word Representation for Machine Translation of Morphologically Rich Languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 10)*, pages 148–157, Cambridge, MA, October.

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)*, pages 311–318, Philadelphia, PA, July.

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

Drahomíra "Johanka" Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-supervised training for the averaged perceptron POS tagger. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 763–771, Athens, Greece, March.

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