Syntax-oriented evaluation measures for machine translation output

Maja Popović and Hermann Ney
RWTH Aachen University
Aachen, Germany
popovic,ney@informatik.rwth-aachen.de

Abstract

We explored novel automatic evaluation measures for machine translation output oriented to the syntactic structure of the sentence: the BLEU score on the detailed Part-of-Speech (POS) tags as well as the precision, recall and F-measure obtained on POS n-grams. We also introduced F-measure based on both word and POS n-grams. Correlations between the new metrics and human judgments were calculated on the data of the first, second and third shared task of the Statistical Machine Translation Workshop. Machine translation outputs in four different European languages were taken into account: English, Spanish, French and German. The results show that the new measures correlate very well with the human judgements and that they are competitive with the widely used BLEU, METEOR and TER metrics.

1 Introduction

We proposed several syntax-oriented automatic evaluation measures based on sequences of POS tags and investigated how they correlate with human judgments. The new measures are the POSBLEU score, i.e. the BLEU score calculated on POS tags instead of words, as well as the POSP, the POSR and the POSF score: precision, recall and F-measure calculated on POS n-grams. In addition to the metrics based only on POS tags, we investigated a WPF score, i.e. an F-measure which takes into account both word and POS n-grams.

The correlations on the document level were computed on the English, French, Spanish and German texts generated by various translation systems in the framework of the first (Koehn and Monz, 2006), second (Callison-Burch et al., 2007) and third shared translation task (Callison-Burch et al., 2008). Preliminary experiments were carried out on the data from the first (2006) and the second task (2007) – Spearman's rank correlation coefficients between the adequacy and fluency scores and the POSBLEU, POSP, POSR and POSF scores were calculated. The POSBLEU and the POSF score were shown to be the most promising, so that these metrics were submitted to the official shared evaluation task 2008. The results of this evaluation showed that these metrics also correlate well on the document level with another human score, i.e. the sentence ranking. However, on the sentence level the results were less promising. The possible reason for this is the main drawback of the metrics based on pure POS tags, i.e. neglecting the lexical aspect. Therefore we also introduced a WPF score which takes into account both word n-grams and POS n-grams.

2 Syntactic-oriented evaluation metrics

We investigated the following metrics oriented on the syntactic structure of a translation output:

- POSBLEU
  The standard BLEU score (Papineni et al., 2002) calculated on POS tags instead of words;
- POSP
  POS n-gram precision: percentage of POS n-grams in the hypothesis which have a counterpart in the reference;
- POSR
  Recall measure based on POS n-grams: percentage of POS n-grams in the reference which are also present in the hypothesis;
- POSF
  POS n-gram based F-measure: takes into account all POS n-grams which have a counter-

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 29–32,
Athens, Greece, 30 March – 31 March 2009.

part, both in the reference and in the hypothesis.

- WPF
  F-measure based both on word and POS n-grams: takes into account all word n-grams and all POS n-grams which have a counterpart both in the corresponding reference and hypothesis.

The prerequisite for all 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.).

The n-gram scores as well as the POSBLEU score are based on fourgrams (i.e. the value of maximal n is 4). For the n-gram-based measures, two types of n-gram averaging were investigated: geometric mean and aritmetic mean. Geometric mean is already widely used in the BLEU score, but 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. However, this problem is probably less critical for POS-based metrics because the tag set sizes are much smaller than vocabulary sizes.

3 Correlations between the new metrics and human judgments

The syntax-oriented evaluation metrics were compared with human judgments 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 value of ρ for an automatic metric, the more similar it is to the human metric. Correlation coefficients between human scores and three well-known automatic measures BLEU, METEOR and TER were calculated as well, in order to see how the new metrics perform in comparison with widely used metrics. The scores were calculated for outputs of translation from Spanish, French and German into English and vice versa. English and German POS tags were produced using the TnT tagger (Brants, 2000), Spanish texts were annotated using the FreeLing analyser (Carreras et al., 2004), and French texts using the TreeTagger[1]. In this way, all references and hypotheses were provided with detailed POS tags.

Experiments on 2006 and 2007 test data

The preliminary experiments with the new evaluation metrics were performed on the data from the first two shared tasks in order to investigate Spearman correlation coefficients ρ between POS-based evaluation measures and the human scores adequacy and fluency. The metrics described in Section 2 (except the WPF score) were calculated for all translation outputs. For each new metric, the ρ coefficient with the adequacy and with the fluency score on the document level were calculated. Then the results were summarised by averaging obtained coefficients over all translation outputs, and the average correlations are presented in Table 1.

2006+2007	adequacy	fluency
BLEU	0.590	0.544
METEOR	0.598	0.538
TER	0.496	0.479
POSBLEU	**0.642**	**0.626**
POSF gm	*0.586*	**0.551**
am	*0.584*	**0.570**
POSR gm	0.572	0.576
am	0.542	0.544
POSP gm	0.551	0.481
am	0.531	0.461

Table 1: Average system-level correlations between automatic evaluation measures and adequacy/fluency scores for 2006 and 2007 test data (gm = geometric mean for n-gram averaging, am = arithmetic mean).

Table 1 shows that the new measures have high ρ coefficients both with respect to the adequacy and to the fluency score. The POSBLEU score has the highest correlations, followed by the POSF score. Furthermore, the POSBLEU score has higher correlations than each of the three widely used metrics, and all the new metrics except the POSP have higher correlations than the TER. The POSF correlations with the fluency are higher than those for the standard metrics, and with the adequacy are comparable to those for the METEOR and the BLEU score.

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

Table 2 presents the percentage of the documents for which the particular new metric has higher correlation than BLEU, METEOR or TER. It can be seen that on the majority of the documents the POSBLEU metric outperforms all three standard measures, especially the correlation with the fluency score. The geometric mean POSF shows similar behaviour, having higher correlation than the standard measures in majority of the cases but slightly less often than the POSBLEU. The POSR has higher correlation than the standard measures in 50-70% of cases, and the POSP score has the lowest percentage, 30-60%. It can be also seen that the geometric mean averaging of the n-grams correlates better with the human judgments more often than the artimetic mean.

Experiments on 2008 test data

For the official shared evaluation task in 2008, the human evaluation scores were different – the adequacy and fluency scores were abandoned being rather time consuming and often inconsistent, and the sentence ranking was proposed as one of the human evaluation scores: the manual evaluators were asked to rank translated sentences relative to each other. RWTH participated in this shared task with the two most promising metrics according to the previous experiments, i.e. POSBLEU and POSF, and the detailed results can be found in (Callison-Burch et al., 2008). It was shown that these metrics also correlate very well with the sentence ranking on the document level. However, on the sentence level the performance was much weaker: a percentage of sentence pairs for which the human comparison yields the same result as the comparison using particular automatic metric was not very high. We believe that the main reason for this is the fact that the metrics based only on the POS tags can assign high scores to translations without correct semantic meaning, because they are taking into account only syntactic structure without taking into account the actual words. For example, if the reference translation is "This sentence is correct", a translation output "This tree is high" would have a POS-based matching score of 100%. Therefore we introduced the WPF score – an F-measure metrics which counts both matching POS n-grams and matching word n-grams.

The ρ coefficients for the POSBLEU, POSF and WPF with the sentence ranking averaged over all translation outputs are shown in Table 3. The correlations for several known metrics are shown as well, i.e. for the BLEU, METEOR and TER along with their variants: METEOR-r denotes the variant optimised for ranking, whereas MBLEU and MTER are BLEU and TER computed using the flexible matching as used in METEOR. It can be seen that the correlation coefficients for all three syntactic metrics are high. The POSBLEU score has the highest correlation with the sentence ranking, followed by POSF and WPF. All three measures have higher average correlation than MTER, MBLEU and BLEU. The purely syntactic metrics outperform also the METEOR scores, whereas the WPF correlations are comparable with those of the METEOR scores.

2008	sentence ranking
BLEU	0.526
MBLEU	0.504
METEOR	0.638
METEOR-r	0.603
MTER	0.318
POSBLEU	0.712
POSF gm	0.663
am	0.661
WPF gm	0.600
am	0.628

Table 3: Average system-level correlations between automatic evaluation measures and human ranking for 2008 test data.

Table 4 presents the percentage of the documents where the particular syntactic metric has higher correlation with the sentence ranking than the particular standard metric. All syntactic metrics have higher correlation than the MTER on almost all documents, and on a large number of documents than the MBLEU score. The correlations for syntactic measures are better than those for the BLEU score for more than 60% of documents. As for the METEOR scores, the syntactic metrics are comparable (about 50%).

4 Conclusions

The results presented in this article suggest that the syntactic information has the potential to strenghten automatic evaluation metrics, and there are many possible directions for future work. We proposed several syntax-oriented evaluation metrics based on the detailed POS tags: the POSBLEU score and POS-n-gram precision, recall and

2006+2007	adequacy			fluency		
	BLEU	METEOR	TER	BLEU	METEOR	TER
POSBLEU	77.3	58.3	75.0	81.8	83.3	83.3
POSF gm	72.7	58.3	75.0	63.6	75.0	83.3
am	68.2	58.3	75.0	63.6	66.7	68.1
POSR gm	63.6	75.0	58.3	68.1	66.7	58.3
am	54.5	75.0	58.3	63.6	58.3	50.0
POSP gm	63.6	50.0	75.0	45.4	50.0	58.3
am	54.5	41.7	66.7	36.4	50.0	58.3

Table 2: Percentage of documents from the 2006 and 2007 shared tasks where the particular new metric has better correlation with adequacy/fluency than the particular standard metric.

2008	BLEU	MBLEU	MTER	METEOR	METEOR-R
POSBLEU	71.4	85.7	92.8	57.1	64.3
POSF am	64.3	78.6	92.8	50.0	50.0
gm	64.3	78.6	92.8	57.1	50.0
WPF am	57.1	64.3	100	42.8	50.0
gm	57.1	64.3	92.8	42.8	50.0

Table 4: Percentage of documents from the 2008 shared task where the new metric has better correlation with the human sentence ranking than the standard metric.

F-measure, i.e. the POSP, POSR, and POSF score. In addition, we introduced a measure which takes into account both POS tags and words: the WPF score. We carried out an extensive analysis of the Spearman's rank correlation coefficients between the syntactic evaluation metrics and the human judgments. The obtained results showed that the new metrics correlate well with human judgments, namely the adequacy and fluency scores, as well as the sentence ranking. The results also showed that the syntax-oriented metrics are competitive with the widely used evaluation measures BLEU, METEOR and TER. Especially promising are the POSBLEU and the POSF score. The correlations of the WPF score are slightly lower than those of the purely POS based metrics – however, this metric has advantage of taking both syntactic and lexical aspect into account.

Acknowledgments

This work was realised as part of the Quaero Programme, funded by OSEO, French State agency for innovati on.

References

Thorsten Brants. 2000. Tnt – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP)*, pages 224–231, Seattle, WA.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-)Evaluation of Machine Translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio, June.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings 4th International Conference on Language Resources and Evaluation (LREC)*, pages 239–242, Lisbon, Portugal, May.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, New York City, June.

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