Automatic Machine Translation Evaluation with Part-of-Speech Information

Aaron L.-F. Han, Derek F. Wong, Lidia S. Chao, and Liangye He

University of Macau, Department of Computer and Information Science
Av. Padre Toms Pereira Taipa, Macau, China
{hanlifengaaron,wutianshui0515}@gmail.com,
{derekfw,lidiasc}@umac.mo

Abstract. One problem of automatic translation is the evaluation of the result. The result should be as close to a human reference translation as possible, but varying word order or synonyms have to be taken into account for the evaluation of the similarity of both. In the conventional methods, researchers tend to employ many resources such as the synonyms vocabulary, paraphrasing, and text entailment data, etc. To make the evaluation model both accurate and concise, this paper explores the evaluation only using Part-of-Speech information of the words, which means the method is based only on the consilience of the POS strings of the hypothesis translation and reference. In this developed method, the POS also acts as the similar function with the synonyms in addition to its syntactic or morphological behaviour of the lexical item in question. Measures for the similarity between machine translation and human reference are dependent on the language pair since the word order or the number of synonyms may vary, for instance. This new measure solves this problem to a certain extent by introducing weights to different sources of information. The experiment results on English, German and French languages correlate on average better with the human reference than some existing measures, such as BLEU, AMBER and MP4IBM1.

Keywords: Natural language processing, Machine translation evaluation, Part-of-Speech, Reference translation.

1 Introduction

With the rapid development of Machine Translation systems, how to evaluate each MT system's quality and what should be the criteria have become the new challenges in front of MT researchers. The commonly used automatic evaluation metrics include the word error rate WER [2], BLEU [3] (the geometric mean of n-gram precision by the system output with respect to reference translations), and NIST [4]. Recently, many other methods were proposed to revise or improve the previous works.

METEOR [5] metric conducts a flexible matching, considering stems, synonyms and paraphrases, which method and formula for computing a score is much more complicated than BLEU's [1]. The matching process involves computationally expensive word alignment. There are some parameters such as the relative weight of recall to precision, the weight for stemming or synonym that should be tuned. Snover [6] discussed that one

I. Habernal and V. Matousek (Eds.): TSD 2013, LNAI 8082, pp. 121–128, 2013.

disadvantage of the Levenshtein distance was that mismatches in word order required the deletion and re-insertion of the misplaced words. They proposed TER by adding an editing step that allows the movement of word sequences from one part of the output to another. AMBER [7] including AMBER-TI and AMBER-NL declare a modified version of BLEU and attaches more kinds of penalty coefficients, combining the n-gram precision and recall with the arithmetic average of F-measure. F15 [8] and F15G3 perform evaluation with the F1 measure (assigning the same weight on precision and recall) over target features as a metric for evaluating translation quality. The target features they defined include TP (be the true positive), TN (the true negative), FP (the false positive), and FN (the false negative rates), etc. To consider the surrounding phrases for a missing token in the translation they employed the gapped word sequence kernels [9] approach to evaluate translations. Other related works include [10], [11] and [12] about the discussion of word order, ROSE [13], MPF and WMPF [14] about the employing of POS information, MP4IBM1 [15] without relying on reference translations, etc.

The evaluation methods proposed previously tend to rely on too many linguistic features (difficult in replicability) or no linguistic information (leading the metrics result in low correlation with human judgments). To address this problem, this paper explores the performance of a novel method only using the consilience of the POS strings of the hypothesis translation and reference translation. This ensures that the linguistic information is considered in the evaluation but it is a very concise model.

2 Linguistic Features

As discussed above, language variability results in no single correct translation and different languages do not always express the same content in the same way. To address the variability phenomenon, researchers used to employ the synonyms, paraphrasing or text entailment as auxiliary information. All of these approaches have their advantages and weaknesses, e.g. the synonyms are difficult to cover all the acceptable expressions. Instead, in the designed metric, we use the part-of-speech (POS) information (also applied by ROSE [13], MPF and WMPF [14]). If the translation sentence of system outputs is a good translation then there is a potential that the output sentence has a similar semantic information with the reference sentence (the two sentences may not contain exactly the same words but with the words that have similar semantic meaning). For example, “there is a big bag” and “there is a large bag” could be the same expression since “big” and “large” has the similar meaning (with POS as adjective). To try this approach, we conduct the evaluation on the POS of the words instead of the words themselves and we do not use other external resources such as synonym dictionaries. We also test the approach by calculating the correlation score of this method with human judgments in the experiment. Assume that we have two sentences: one reference and one system output translation. Firstly, we extract the POS of each word. Then, we calculate the similarity of these two sentences through the alignment of their POS information.

3 Calculation Methods

3.1 Design of hLEPOR Metric

First, we introduce the mathematical harmonic mean for multi-variables (n variables $(X_1, X_2, \ldots, X_n)$).

$$Harmonic(X_1, X_2, ..., X_n) = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}} \tag{1}$$

where n means the number of variables (also named as factors). Then, the weighted harmonic mean for multi-variables is:

$$Harmonic(w_{X_1}X_1, w_{X_2}X_2, ..., w_{X_n}X_n) = \frac{\sum_{i=1}^{n} w_{X_i}}{\sum_{i=1}^{n} \frac{w_{X_i}}{X_i}} \tag{2}$$

where w_{X_i} presents the weight assigned to the corresponding variable X_i. Finally, the proposed evaluation metric $hLEPOR$ (tunable Harmonic mean of Length Penalty, Precision, n-gram Position difference Penalty and Recall) is designed as:

$$hLEPOR = Harmonic(w_{LP}LP, w_{NPosPenal}NPosPenal, w_{HPR}HPR) \tag{3}$$

$$= \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{Factor_i}} = \frac{w_{LP} + w_{NPosPenal} + w_{HPR}}{\frac{w_{LP}}{LP} + \frac{w_{NPosPenal}}{NPosPenal} + \frac{w_{HPR}}{HPR}}$$

where LP, $NPosPenal$ and HPR are three factors in $hLEPOR$ and will be introduced in the following. Three tunable weights parameters w_{LP}, $w_{NPosPenal}$ and w_{HPR} are assigned to the three factors respectively.

3.2 Design of Internal Factors

Length Penalty. In the Eq. (3), LP means Length penalty to embrace the penalty for both longer and shorter system outputs compared with the reference translations:

$$LP = \begin{cases} e^{1-\frac{r}{c}} & : \quad c < r \\ 1 & : \quad c = r \\ e^{1-\frac{c}{r}} & : \quad c > r \end{cases} \tag{4}$$

where c and r mean the sentence length of candidate translation and reference translation respectively.

N-gram Position Difference Penalty. In the Eq.(3), the $NPosPenal$ is defined as:

$$NPosPenal = e^{-NPD} \tag{5}$$

where NPD means n-gram position difference penalty. The $NPosPenal$ value is designed to compare the POS order in the sentences between reference translation and output translation. The NPD is defined as:

$$NPD = \frac{1}{Length_{output}} \sum_{i=1}^{Length_{output}} |PD_i| \tag{6}$$

where $Length_{output}$ represents the length of system output sentence and PD_i means the n-gram position difference value of aligned POS between output and reference sentences. Every POS from both output translation and reference should be aligned only once (one-to-one alignment). When there is no match, the value of PD_i will be zero as default for this output POS.

To calculate the NPD value, there are two steps: aligning and calculating. To begin with, the context-dependent n-gram alignment task: we use the n-gram method and assign higher priority on it, which means we take into account the surrounding context (surrounding POS) of the potential POS to select a better matching pairs between the output and the reference. If there are both nearby matching or there is no matched POS around the potential pairs, then we consider the nearest matching to align as a backup choice. The alignment direction is from output sentence to the references.

See example in Figure 1. In the second step (calculating step), we label each POS with its position number divided by the corresponding sentence length for normalization, and then using the Eq. (6) to finish the calculation.

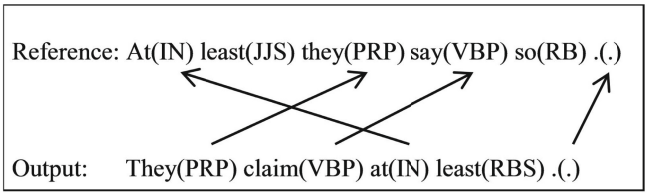

Fig. 1. Example of n-gram POS alignment

We also use the example in Figure 1 for the NPD introduction (Figure 2). In the example, when we label the position number of output sentence we divide the numerical position (from 1 to 5) of the current POS by the sentence length 5. For the reference sentence it is the similar step. After we get the NPD value, using the Eq. (5), the values of $NPosPenal$ are calculated.

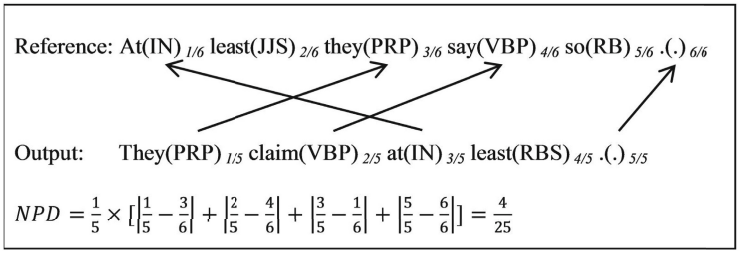

Fig. 2. Example of NPD calculation

Precision and Recall. Precision is designed to reflect the accurate rate of outputs while recall means the loyalty to the references. In the Eq. (3), HPR means the weighted Harmonic mean of precision and recall i.e. $Harmonic(\alpha R, \beta P)$, with parameters α and β as the tunable weights for recall and precision respectively.

$$Harmonic(\alpha R, \beta P) = \frac{\alpha + \beta}{\frac{\alpha}{R} + \frac{\beta}{P}} \tag{7}$$

$$P = \frac{aligned_{num}}{system_{length}} \tag{8}$$

$$R = \frac{aligned_{num}}{reference_{length}} \tag{9}$$

where $aligned_{num}$ represents the number of successfully aligned (matched) POS appearing both in translation and reference, $system_{length}$ and $reference_{length}$ specify the sentence length of system output and reference respectively.

System-level hLEPOR. We have introduced the calculation of $hLEPOR$ on single output sentence, and we should consider a proper way to calculate the value when the cases turn into document (or system) level. We perform the system-level $hLEPOR$ as below.

$$hLEPOR_{sys} = Harmonic(w_{LP}LP_{sys}, w_{NPosPenal}PosPenalty_{sys}, w_{HPR}HPR_{sys}) \tag{10}$$

As shown in the formula, to calculate the system-level score $hLEPOR_{sys}$, we should firstly calculate the system-level scores of its factors LP_{sys}, $PosPenalty_{sys}$ and HPR_{sys}. The system level factor scores are calculated by the arithmetic means of the corresponding sentence-level factor scores.

4 Experiments

We trained $hLEPOR$ and tuned the parameters on the public ACL WMT 2008[1] data. There are five languages in the WMT 2008 data including English, Spanish, German, French and Czech; however, we currently did not find proper parser tools for the Spanish and Czech languages. So we tested on the English, German and French languages using the Berkeley parsers [16] to extract the POS information of the tested sentences. Thus, there are four language pairs in our tested corpora: from German and French to English, and the inverse. The parameter values on all language pairs are shown in Table 1.

The tested corpora we used are from ACL WMT 2011[2]. There are more than one hundred MT systems offering their output translation results, with most MT systems

[1] http://www.statmt.org/wmt08/
[2] http://www.statmt.org/wmt11/

Table 1. Values of tuned parameters

Parameters		
(α, β)	n-gram POS Alignment	Weights(HPR:LP:NPosPenal)
(9,1)	2-gram	3:2:1

statistical-based except for five rule-based ones. The gold standard reference data for those corpora consists of 3003 sentences. For each language pairs, there are different numbers of participated MT systems. Automatic MT evaluation systems are differed by calculating their Spearman rank correlation coefficient with the human judgment results [17].

Table 2. Correlation coefficients with human judgments

	Correlation Score with Human Judgment				
	Other-to-English		English-to-Other		
Metrics	**DE-EN**	**FR-EN**	**EN-DE**	**EN-FR**	**Mean score**
$hLEPOR$	0.83	0.74	0.84	0.82	**0.81**
MPF	0.69	0.87	0.63	0.89	0.77
WMPF	0.66	0.87	0.61	0.89	0.76
AMBER-TI	0.63	0.94	0.54	0.84	0.74
AMBER	0.59	0.95	0.53	0.84	0.73
AMBER-NL	0.58	0.94	0.45	0.83	0.7
METEOR-1.3	0.71	0.93	0.3	0.85	0.70
ROSE	0.59	0.86	0.41	0.86	0.68
BLEU	0.48	0.85	0.44	0.86	0.66
F15G3	0.48	0.88	0.3	0.84	0.63
F15	0.45	0.87	0.19	0.85	0.59
MP4IBM1	0.56	0.08	0.91	0.61	0.54
TER	0.33	0.77	0.12	0.84	0.52

We compare the experiments results with several classic metrics including BLEU, METEOR, TER and some latest ones (e.g. MPF, ROSE, F15, AMBER, MP4IBM1). The system level correlation coefficients of these metrics with human judgments are shown in the Table 2 which is ranked by the mean correlation scores of the metrics on four language pairs. Several conclusions from the results could be drawn: first, many evaluation metrics performed well in certain language pairs but weak on others, e.g. WMPF results in 0.89 correlation with human judgments on English-to-French corpus but down to 0.61 score on English-to-German, F15 gets 0.87 score on French-to-English but 0.45 on German-to-English, ROSE performs well on both French-to-English and English-to-French but worse on Germen-to-English and English-to-German. Second, recently proposed evaluation metrics (e.g. MPF and AMBER) generally perform better than the traditional ones (e.g. BLEU and TER), showing an improvement of the research work.

5 Conclusion and Perspectives

To make the evaluation model both accurate and concise, instead of using the synonyms vocabularies, paraphrasing and text entailment that are commonly used by other researchers, this paper explores the use of the POS information of the words sequences. What is noticing is that some researchers have used the n-gram method on the words alignment (e.g. BLEU using 1-gram to 4-gram), other researchers used the POS information in the similarity calculation by counting the number of corresponding POS. However, this paper employs the n-gram method on the POS alignment. Since this developed metric only relies on the consilience of the POS strings of the evaluated sentence even without using the surface words, it has a potential to be further developed as a reference independent metric. The main difference of this paper and our previous work LEPOR (the product of factors, perform on words) [18] is that this method groups the factors based on mathematical weighted harmonic mean, instead of the simple product of factors, and this method is performed on the POS instead of words. The overall weighted harmonic mean allows to tune the model neatly according to different circumstances, and the POS information can act as part of synonyms in addition to the syntactic or morphological behaviour of the lexical item.

Even though the designed metric has shown promising performances on the tested language pairs (EN to DE and FR, and the inverse direction). There are several weaknesses of this metric. Firstly, the POS codes are not language independent, so this method may not work well on the distant languages e.g. English and Japanese. Secondly, the parsing accuracy will effect the performance of the evaluation. Thirdly, to make the evaluation model concise, this work only uses the POS information without considering the surface words. To address these weaknesses, in the future work, more language pairs will be tested, other POS generation tools will be explored, and the combination of surface words and POS will be employed.

Acknowledgments. The authors wish to thank the anonymous reviewers for many helpful comments.

References

1. Koehn, P.: Statistical Machine Translation (University of Edinburgh). Cambridge University Press (2010)
2. Su, K.-Y., Wu, M.-W., Chang, J.-S.: A New Quantitative Quality Measure for Machine Translation Systems. In: Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France, pp. 433–439 (July 1992)
3. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the ACL 2002, Philadelphia, PA, USA, pp. 311–318 (2002)
4. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research, San Diego, California, USA, pp. 138–145 (2002)
5. Banerjee, S., Lavie, A.: Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of ACL-WMT, Prague, Czech Republic, pp. 65–72 (2005)

6. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of the Conference of the Association for Machine Translation in the Americas, Boston, USA, pp. 223–231 (2006)
7. Chen, B., Kuhn, R.: Amber: A modified bleu, enhanced ranking metric. In: Proceedings of ACL-WMT, Edinburgh, Scotland, UK, pp. 71–77 (2011)
8. Bicici, E., Yuret, D.: RegMT system for machine translation, system combination, and evaluation. In: Proceedings ACL-WMT, Edinburgh, Scotland, UK, pp. 323–329 (2011)
9. Taylor, J.S., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press (2004)
10. Wong, B.T.-M., Kit, C.: Word choice and word position for automatic MT evaluation. In: Workshop: MetricsMATR of the Association for Machine Translation in the Americas, Waikiki, Hawai, USA, 3 pages (2008)
11. Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H.: Automatic evaluation of translation quality for distant language pairs. In: Proceedings of the 2010 Conference on EMNLP, Cambridge, MA, pp. 944–952 (2010)
12. Talbot, D., Kazawa, H., Ichikawa, H., Katz-Brown, J., Seno, M., Och, F.: A Lightweight Evaluation Framework for Machine Translation Reordering. In: Proceedings of the Sixth ACL-WMT, Edinburgh, Scotland, UK, pp. 12–21 (2011)
13. Song, X., Cohn, T.: Regression and ranking based optimisation for sentence level MT evaluation. In: Proceedings of the ACL-WMT, Edinburgh, Scotland, UK, pp. 123–129 (2011)
14. Popovic, M.: Morphemes and POS tags for n-gram based evaluation metrics. In: Proceedings of ACL-WMT, Edinburgh, Scotland, UK, pp. 104–107 (2011)
15. Popovic, M., Vilar, D., Avramidis, E., Burchardt, A.: Evaluation without references: IBM1 scores as evaluation metrics. In: Proceedings of the ACL-WMT, Edinburgh, Scotland, UK, pp. 99–103 (2011)
16. Petrov, S., Barrett, L., Thibaux, R., Klein, D.: Learning accurate, compact, and interpretable tree annotation. In: Proceedings of the 21st ACL, Sydney, pp. 433–440 (July 2006)
17. Callison-Bruch, C., Koehn, P., Monz, C., Zaidan, O.F.: Findings of the 2011 Workshop on Statistical Machine Translation. In: Proceedings of ACL-WMT, Edinburgh, Scotland, UK, pp. 22–64 (2011)
18. Han, A.L.-F., Wong, D.F., Chao, L.S.: LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012): Posters, Mumbai, India, pp. 441–450 (2012)