Language-independent Model for Machine Translation Evaluation with Reinforced Factors

Aaron Li-Feng Han, Derek F. Wong, Lidia S. Chao, Liangye He
Yi Lu, Junwen Xing, and **Xiaodong Zeng**
Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory
Department of Computer and Information Science
University of Macau, Macau S.A.R., China
hanlifengaaron@gmail.com, {derekfw, lidiasc}@umac.mo
{wutianshui0515,takamachi660,nlp2ct.anson,nlp2ct.samuel}@gmail.com

Abstract

The conventional machine translation evaluation metrics tend to perform well on certain language pairs but weak on other language pairs. Furthermore, some evaluation metrics could only work on certain language pairs not language-independent. Finally, no considering of linguistic information usually leads the metrics result in low correlation with human judgments while too many linguistic features or external resources make the metrics complicated and difficult in replicability. To address these problems, a novel language-independent evaluation metric is proposed in this work with enhanced factors and optional linguistic information (part-of-speech, n-grammar) but not very much. To make the metric perform well on different language pairs, extensive factors are designed to reflect the translation quality and the assigned parameter weights are tunable according to the special characteristics of focused language pairs. Experiments show that this novel evaluation metric yields better performances compared with several classic evaluation metrics (including BLEU, TER and METEOR) and two state-of-the-art ones including ROSE and MPF.

1 Introduction

The machine translation (MT) began as early as in the 1950s (Weaver, 1955) and gained a big progress science the 1990s due to the development of computers (storage capacity and computational power) and the enlarged bilingual corpora (Marino et al., 2006), e.g. (Och, 2003) presented MERT (Minimum Error Rate Training) for log-linear statistical machine translation (SMT) models to achieve better translation quality, (Su et al., 2009) used the Thematic Role Templates model to improve the translation and (Xiong et al., 2011) employed the maximum-entropy model etc. The statistical MT (Koehn, 2010) became mainly approaches in MT literature. Due to the wide-spread development of MT systems, the MT evaluation becomes more and more important to tell us how well the MT systems perform and whether they make some progress. However, the MT evaluation is difficult because some reasons, e.g. language variability results in no single correct translation, the natural languages are highly ambiguous and different languages do not always express the same content in the same way (Arnold, 2003).

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 earliest human assessment methods include the intelligibility (measuring how understandable the sentence is) and fidelity (measuring how much information the translated sentence retains compared to the original) used by the Automatic Language Processing Advisory Committee (ALPAC) around 1966 (Carroll, 1966), and the afterwards proposed adequacy (similar as fidelity), fluency (whether the sentence is well-formed and fluent) and comprehension (improved intelligibility) by Defense Advanced Research Projects Agency (DARPA) of US (White et al., 1994). The manual evaluations suffer the main disadvantage that it is time-consuming and thus too expensive to do frequently.

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 215–222.

The early automatic evaluation metrics include the word error rate WER (Su et al., 1992) (edit distance between the system output and the closest reference translation), position independent word error rate PER (Tillmann et al., 1997) (variant of WER that disregards word ordering), BLEU (Papineni et al., 2002) (the geometric mean of n-gram precision by the system output with respect to reference translations), NIST (Doddington, 2002) (adding the information weight) and GTM (Turian et al., 2003). Recently, many other methods were proposed to revise or improve the previous works.

One of the categories is the lexical similarity based metric. The metrics of this kind include the edit distance based method, such as the TER (Snover et al., 2006) and the work of (Akiba et al., 2001) in addition to WER and PER, the precision based method such as SIA (Liu and Gildea, 2006) in addition to BLEU and NIST, recall based method such as ROUGE (Lin and Hovy, 2003), the word order information utilized by (Wong and Kit, 2008), (Isozaki et al., 2010) and (Talbot et al., 2011), and the combination of precision and recall such as Meteor-1.3 (Denkowski and Lavie, 2011) (an modified version of Meteor, includes ranking and adequacy versions and has overcome some weaknesses of previous version such as noise in the paraphrase matching, lack of punctuation handling and discrimination between word types), BLANC (Lita et al., 2005), LEPOR (Han et al., 2012) and PORT (Chen et al., 2012). Another category is the employing of linguistic features. The metrics of this kind include the syntactic similarity such as the Part-of-Speech information used by ROSE (Song and Cohn, 2011) and MPF (Popovic, 2011), and phrase information employed by (Echizen-ya and Araki, 2010) and (Han et al., 2013b); the semantic similarity such as Textual entailment used by (Mirkin et al., 2009), Synonyms by (Chan and Ng, 2008), paraphrase by (Snover et al., 2009).

The evaluation methods proposed previously suffer from several main weaknesses more or less: perform well in certain language pairs but weak on others, which we call the language-bias problem; consider no linguistic information (not reasonable from the aspect of linguistic analysis) or too many linguistic features (making it difficult in replicability), which we call the extremism problem; present incomprehensive factors (e.g. BLEU focus on precision only). To address these problems, a novel automatic evaluation metric is proposed in this paper with enhanced factors, tunable parameters and optional linguistic information (part-of-speech, n-gram).

2 Designed Model

2.1 Employed Internal Factors

Firstly, we introduce the internal factors utilized in the calculation model.

2.1.1 Enhanced Length Penalty

Enhanced length penalty ELP is designed to put the penalty on both longer and shorter system output translations (an enhanced version of the brevity penalty in BLEU):

$$ELP = \begin{cases} e^{1-\frac{r}{c}} & : \ c < r \\ e^{1-\frac{c}{r}} & : \ c \geq r \end{cases} \quad (1)$$

where the parameters c and r are the sentence length of automatically output (candidate) and reference translation respectively.

2.1.2 N-gram Position Difference Penalty

The N-gram Position Difference Penalty $NPosPenal$ is developed to compare the word order between the output and reference translation.

$$NPosPenal = e^{-NPD} \quad (2)$$

where NPD is defined as:

$$NPD = \frac{1}{Length_{output}} \sum_{i=1}^{Length_{output}} |PD_i| \quad (3)$$

where $Length_{output}$ is the length of system output sentence and PD_i means the position difference value of each output word. Every word from both output translation and reference should be aligned only once. When there is no match, the value of PD_i is assigned with zero as default for this output token.

Two steps are designed to measure the NPD value. The first step is the context-dependent n-gram alignment: we use the n-gram method and assign it with higher priority, which means the surrounding context of the potential words are considered when selecting the matched pairs between the output and reference sentence. The nearest match is accepted as a backup choice to establish the alignment, if there are both nearby matching or there is no other matched words surrounding the potential word pairs. The one-direction alignment is from output sentence to the reference.

Assuming that w_x represents the current word in output sentence and w_{x+k} means the kth word to the previous ($k < 0$) or following ($k > 0$). On the other hand, w_y^r means the word matching w_x in the references, and w_{y+j}^r has the similar meaning as w_{x+k} but in reference sentence. The variable $Distance$ is the position difference value between the matching word in outputs and references. The operation process and pseudo code of the context-dependent n-gram word alignment algorithm are shown in Figure 1 (with → as the alignment). There is an example in Figure 2. In the calculating step, each word is labeled with the quotient value of its position number divided by sentence length (the total number of the tokens in the sentence).

Let's see the example in Figure 2 for the NPD introduction (Figure 3). Each output word is labeled with the position quotient value from $1/6$ to $6/6$ (indicating the word position marked by sentence length which is 6). The words in the reference sentence is labeled using the same subscripts.

2.1.3 Precision and Recall

Precision and recall are two commonly used criteria in the NLP literature. We use the HPR to represent the weighted Harmonic mean of precision and recall, i.e. $Harmonic(\alpha R, \beta P)$. The weights are the tunable parameters α and β.

$$HPR = \frac{(\alpha+\beta)Precision \times Recall}{\alpha Precision + \beta Recall} \quad (4)$$

$$Precision = \frac{Aligned_{num}}{Length_{output}} \quad (5)$$

$$Recall = \frac{Aligned_{num}}{Length_{reference}} \quad (6)$$

where $Aligned_{num}$ represents the number of successfully matched words appearing both in translation and reference.

2.2 Sentence Level Score

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

$$Harmonic(X_1, X_2, ..., X_n) = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}} \quad (7)$$

where n is the number of factors. Then, the weighted harmonic mean for multi-variables is:

$$Harmonic(w_{X_1}X_1, w_{X_2}X_2, ..., w_{X_n}X_n) =$$

$$\frac{\sum_{i=1}^{n} w_{X_i}}{\sum_{i=1}^{n} \frac{w_{X_i}}{X_i}} \quad (8)$$

where w_{X_i} is the weight of variable X_i. Finally, the sentence level score of the developed evaluation metric $hLEPOR$ (Harmonic mean of enhanced Length Penalty, Precision, n-gram Position difference Penalty and Recall) is measured by:

$$hLEPOR =$$

$$= \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{Factor_i}} = \frac{w_{ELP} + w_{NPosPenal} + w_{HPR}}{\frac{w_{ELP}}{ELP} + \frac{w_{NPosPenal}}{NPosPenal} + \frac{w_{HPR}}{HPR}} \quad (9)$$

where ELP, $NPosPenal$ and HPR are the three factors explained in previous section with tunable weights w_{ELP}, $w_{NPosPenal}$ and w_{HPR} respectively.

2.3 System-level Score

The system level score is the arithmetical mean of the sentence scores as below.

$$\overline{hLEPOR} = \frac{1}{SentNum} \sum_{i=1}^{SentNum} hLEPOR_i \quad (10)$$

where $\overline{hLEPOR}$ represents the system-level score of $hLEPOR$, $SentNum$ specifies the sentence number of the test document, and $hLEPOR_i$ means the score of the ith sentence.

3 Enhanced Version

This section introduces an enhanced version of the developed metric $hLEPOR$ as $hLEPOR_E$. As discussed by many researchers, language variability results in no single correct translation and different languages do not always express the same content in the same way. In addition to the augmented factors of the designed metric $hLEPOR$, we present that optional linguistic information can be combined into this metric concisely. As an example, we will show how the part-of-speech (POS) information can be employed into this metric. First, we calculate the system-level $hLEPOR$ scores on the surface words ($\overline{hLEPOR_{word}}$). Then we employ the same algorithms of $hLEPOR$ on the corresponding POS sequences of the words ($\overline{hLEPOR_{POS}}$). Finally, we combine this two system-level scores together with tunable weights (w_{hw} and w_{hp}) as the final score.

$$\overline{hLEPOR_E} = \frac{1}{w_{hw} + w_{hp}}(w_{hw}\overline{hLEPOR_{word}} + w_{hp}\overline{hLEPOR_{POS}}) \quad (11)$$

We mention the POS information because it sometimes acts as the similar function with the synonyms, e.g. "there is a big bag" and "there is a large bag" could be the same meaning but with different surface words "big" and "large" (the same POS adjective). The POS information has been proved helpful in the research works of ROSE (Song and Cohn, 2011) and MPF (Popovic, 2011). The POS information could be replaced by any other concise linguistic information in our designed model.

4 Evaluating the Evaluation Metric

In order to distinguish the reliability of different MT evaluation metrics, Spearman rank correlation coefficient ρ is commonly used to calculate the correlation in the annual workshop of statistical machine translation (WMT) for Association of Computational Linguistics (ACL) (Callison-Burch et al., 2011). When there are no ties, Spearman rank correlation coefficient is calculated as:

$$\rho_{\phi(XY)} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad (12)$$

where d_i is the difference-value (D-value) between the two corresponding rank variables $\vec{X} = \{x_1, x_2, ..., x_n\}$ and $\vec{Y} = \{y_1, y_2, ..., y_n\}$ describing the system ϕ and n is the number of variables in the system.

5 Experiments

The experiment corpora are from the ACL's special interest group of machine translation SIGMT (WMT workshop) which contain eight corpora including English-to-other (Spanish, Czech, French and German) and other-to-English. There are indeed a lot of linguistic POS tagger tools for different languages available. We conduct an evaluation with different POS taggers, and find that the employing of POS information can make an increase of the correlation score with human judgment for some language pairs but little or no effect on others. The employed POS tagging tools include Berkeley POS tagger for French, English and German (Petrov et al., 2006), COMPOST Czech morphology tagger (Collins, 2002) and TreeTagger Spanish tagger (Schmid, 1994). To avoid the overfitting problem, the WMT 2008[1] data are used in the development stage for the tuning of the parameters and the WMT 2011 corpora are used in testing. The tuned parameter values for different language pairs are shown in Table 1. The abbreviations EN, CZ, DE, ES and FR mean English, Czech, German, Spanish and French respectively. In the n-gram word (POS) alignment, bigram is selected in all the language pairs. To make the model concise using as fewer of external resources as possible, the value of "N/A" means the POS information of that language pair is not employed due to that it makes little or no effect in the correlation scores. The label "(W)" and "(POS)" means the parameters tuned on word and POS respectively. The "NPP" means $NPosPenal$ to save window space. The tuned parameter values also prove that different language pairs embrace different characteristics.

The testing results on WMT 2011[2] corpora are shown in Table 2. The comparisons with language-independent evaluation metrics include the classic metrics (BLEU, TER and METEOR) and two state-of-the-art metrics MPF and ROSE. We select MPF and ROSE because that these two metrics also employ the POS information and MPF yielded the highest correlation score with human judgments among all the language-independent metrics (performing on eight language pairs) in WMT 2011. The numbers of participated automatic MT systems in WMT 2011 are 10, 22, 15 and 17 respectively for English-to-other (CZ, DE, ES and FR) and 8, 20, 15 and 18 respectively for the opposite translation direction. The gold standard reference data for those corpora consists of 3,003 sentences offered by manual work. Automatic MT e-

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

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

Ratio	**Other-to-English**				**English-to-Other**			
	CZ-EN	DE-EN	ES-EN	FR-EN	EN-CZ	EN-DE	EN-ES	EN-FR
HPR:ELP:NPP(W)	7:2:1	3:2:1	7:2:1	3:2:1	3:2:1	1:3:7	3:2:1	3:2:1
HPR:ELP:NPP(POS)	N/A	3:2:1	N/A	3:2:1	N/A	7:2:1	N/A	3:2:1
$\alpha:\beta$(W)	1:9	9:1	1:9	9:1	9:1	9:1	9:1	9:1
$\alpha:\beta$(POS)	N/A	9:1	N/A	9:1	N/A	9:1	N/A	9:1
$w_{hw}:w_{hp}$	N/A	1:9	N/A	9:1	N/A	1:9	N/A	9:1

Table 1: **Values of tuned weight parameters**

valuation metrics are evaluated by the correlation coefficient with the human judgments.

Several conclusions could be drawn from the results. First, some evaluation metrics show good performances on part of the language pairs but low performances on others, e.g ROSE results in 0.92 correlation with human judgments on Spanish-to-English corpus but down to 0.41 score on English-to-German; METEOR gets 0.93 score on French-to-English but 0.3 on English-to-German. Second, $hLEPOR_E$ generally yields good performances on different language pairs except for the English-to-Czech and results in the highest Mean correlation score 0.83 on eight corpora. Third, the recently developed methods (e.g. MPF, 0.81 mean score) correlate better with human judgments than the traditional ones (e.g. BLEU, 0.74 means score), indicating an improvement of the researches. Finally, no metric can yield high performance on all the language pairs, which shows that there remains large potential to achieve improvement.

6 Conclusion and Perspectives

This work proposes a language-independent model for machine translation evaluation. Considering the different characteristics of different languages, $hLEPOR_E$ has been extensively designed from different aspects. That spans from word order (context-dependent n-gram alignment), output accuracy (precision), and loyalty (recall) to translation length performance (sentence length). D-ifferent weight parameters are assigned to adjust the importance of each factor, for instance, the word position could be free in some languages but strictly constrained in other languages. In practice, these employed features by $hLEPOR_E$ are also the vital ones when people facilitate language translation. This is the philology behind the formulation and the study of this work, and we believe human's translation ideology is the exact direction that MT systems should try to approach. Furthermore, this work specifies that different external resources or linguistic information could be integrated into this model easily. As suggested by other works, e.g. (Avramidis et al., 2011), the POS information is considered in the experiments and shows some improvements on certain language pairs.

There are several main contributions of this paper compared with our previous work (Han et al., 2013). This work combines the utilizing of surface words and linguistic features together (instead of relying on the consilience of the POS sequence only). This paper measures the system-level $hLEPOR$ score by the arithmetical mean of each sentence-level score (instead of the Harmonic mean of system-level internal factors). This paper shows the performances of enhanced method $hLEPOR_E$ on all the eight language pairs released by WMT official web (instead of part language pairs by previous work) and most of the performances have achieved improvements than previous work on the same language pairs (e.g. the correlation score on German-English is 0.86 increased from 0.83; the correlation score on French-English is 0.92 increased from 0.74.). Other potential linguistic features are easily to be employed into the flexible model built in this paper.

There are also several aspects that should be addressed in the future works. Firstly, more language pairs, in addition to the European languages, will be tested such as Japanese, Korean and Chinese and the performances of linguistic features (e.g. POS tagging) will also be explored on the new language pairs. Secondly, the tuning of weight parameters to achieve high correlation with human judgments during the development period will be automatically performed. Thirdly, since the use of multiple references helps the usual translation

Metrics	Other-to-English				English-to-Other				Mean
	CZ-EN	DE-EN	ES-EN	FR-EN	EN-CZ	EN-DE	EN-ES	EN-FR	
$hLEPOR_E$	0.93	0.86	0.88	0.92	0.56	0.82	0.85	0.83	**0.83**
MPF	0.95	0.69	0.83	0.87	0.72	0.63	0.87	0.89	0.81
ROSE	0.88	0.59	0.92	0.86	0.65	0.41	0.9	0.86	0.76
METEOR	0.93	0.71	0.91	0.93	0.65	0.3	0.74	0.85	0.75
BLEU	0.88	0.48	0.9	0.85	0.65	0.44	0.87	0.86	0.74
TER	0.83	0.33	0.89	0.77	0.5	0.12	0.81	0.84	0.64

Table 2: **Correlation coefficients with human judgments**

quality measures correlate with the human judging, the scheme of how to use the multiple references will be designed.

Acknowledgments.

The authors are grateful to the Science and Technology Development Fund of Macau and the Research Committee of the University of Macau for the funding support for our research, under the reference No. 017/2009/A and RG060/09-10S/CS/FST. The authors also wish to thank the anonymous reviewers for many helpful comments.

References

Akiba, Y., K. Imamura, and E. Sumita. 2001. Using Multiple Edit Distances to Automatically Rank Machine Translation Output. *Proceedings of MT Summit VIII* , Santiago de Compostela, Spain.

Arnold, D. 2003. Why translation is difficult for computers. *In Computers and Translation: A translator's guide* , Benjamins Translation Library.

Avramidis, E., Popovic, M., Vilar, D., Burchardt, A. 2011. Evaluate with Confidence Estimation: Machine ranking of translation outputs using grammatical features. *Proceedings of ACL-WMT* , pages 65-70, Edinburgh, Scotland, UK.

Callison-Bruch, C., Koehn, P., Monz, C. and Zaidan, O. F. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. *Proceedings of ACL-WMT* , pages 22-64, Edinburgh, Scotland, UK.

Carroll, J. B. 1966. Aan experiment in evaluating the quality of translation. *Languages and machines: computers in translation and linguistics* , Automatic Language Processing Advisory Committee (ALPAC), Publication 1416, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, page 67-75.

Chan, Y. S. and Ng, H. T. 2008. MAXSIM: A maximum similarity metric for machine translation evaluation. *Proceedings of ACL 2008: HLT* , pages 55–62.

Chen, Boxing, Roland Kuhn and Samuel Larkin. 2012. PORT: a Precision-Order-Recall MT Evaluation Metric for Tuning. *Proceedings of 50th ACL)* , pages 930–939, Jeju, Republic of Korea.

Collins, M. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. *Proceedings of the ACL-02 conference, Volume 10 (EMNLP 02)* , pages 1-8. Stroudsburg, PA, USA .

Denkowski, M. and Lavie, A. 2011. Meteor: Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. *Proceedings of (ACL-WMT)* ,pages 85-91, Edinburgh, Scotland, UK.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the second international conference on Human Language Technology Research* , pages 138-145, San Diego, California, USA.

Echizen-ya, H. and Araki, K. 2010. Automatic evaluation method for machine translation using nounphrase chunking. *Proceedings of ACL 2010* , pages 108–117. Association for Computational Linguistics.

Han, Aaron L.-F., Derek F. Wong, and Lidia S. Chao. 2012. LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors. *Proceedings of the 24th International Conference of COLING*, Posters, pages 441-450, Mumbai, India.

Han, Aaron L.-F., Derek F. Wong, Lidia S. Chao, and Liangye He. 2013. Automatic Machine Translation Evaluation with Part-of-Speech Information. *Proceedings of the 16th International Conference of Text, Speech and Dialogue (TSD 2013)*, LNCS Volume Editors: Vaclav Matousek et al. Springer-Verlag Berlin Heidelberg. Plzen, Czech Republic.

Han, Aaron L.-F., Derek F. Wong, Lidia S. Chao, Liangye He, Shuo Li, and Ling Zhu. 2013b. Phrase Mapping for French and English Treebank and the Application in Machine Translation Evaluation. *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology, (GSCL 2013)*, LNCS Volume Editors: Iryna Gurevych, Chris Biemann and Torsten Zesch. Darmstadt, Germany.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. 2010. Automatic evaluation of translation quality for distant language pairs. *Proceedings of the 2010 Conference on EMNLP* , pages 944–952, Cambridge, MA.

Koehn, P. 2010. Statistical Machine Translation. *Cambridge University Press* .

Marino, B. Jose, Rafael E. Banchs, Josep M. Crego, Adria de Gispert, Patrik Lambert, Jose A. Fonollosa, and Marta R. Costa-jussa. 2006. N-gram based machine translation. *Journal of the Computational Linguistics* ,Vol. 32, No. 4. pp. 527-549, MIT Press.

Lin, Chin-Yew and E.H. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. *Proceedings of HLT-NAACL 2003*, Edmonton, Canada.

Lita, Lucian Vlad, Monica Rogati and Alon Lavie. 2005. BLANC: Learning Evaluation Metrics for MT. *Proceedings of the HLT/EMNLP*, pages 740–747, Vancouver.

Liu D. and Daniel Gildea. 2006. Stochastic iterative alignment for machine translation evaluation. *Proceedings of ACL-06*, Sydney.

Mirkin S., Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. Source-Language Entailment Modeling for Translating Unknown Terms. *Proceedings of the ACL-IJCNLP 2009)* , pages 791–799, Suntec, Singapore.

Och, F. J. 2003. Minimum Error Rate Training for Statistical Machine Translation. *Proceedings of ACL-2003* , pp. 160-167.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the ACL 2002* , pages 311-318, Philadelphia, PA, USA.

Petrov, S., Leon Barrett, Romain Thibaux, and Dan Klein 2006. Learning accurate, compact, and interpretable tree annotation. *Proceedings of the 21st ACL* , pages 433–440, Sydney.

Popovic, M. 2011. Morphemes and POS tags for n-gram based evaluation metrics. *Proceedings of WMT* , pages 104-107, Edinburgh, Scotland, UK.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing* , Manchester, UK.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul J. 2006. A study of translation edit rate with targeted human annotation. *Proceedings of the AMTA*, pages 223-231, Boston, USA.

Snover, Matthew G., Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *J. Machine Translation*, 23: 117-127.

Song, X. and Cohn, T. 2011. Regression and ranking based optimisation for sentence level MT evaluation. *Proceedings of the WMT* , pages 123-129, Edinburgh, Scotland, UK.

Su, Hung-Yu and Chung-Hsien Wu. 2009. Improving Structural Statistical Machine Translation for Sign Language With Small Corpus Using Thematic Role Templates as Translation Memory. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING* , VOL. 17, NO. 7.

Su, Keh-Yih, Wu Ming-Wen and Chang Jing-Shin. 1992. A New Quantitative Quality Measure for Machine Translation Systems. *Proceedings of COLING*, pages 433–439, Nantes, France.

Talbot, D., Kazawa, H., Ichikawa, H., Katz-Brown, J., Seno, M. and Och, F. 2011. A Lightweight Evaluation Framework for Machine Translation Reordering. *Proceedings of the WMT*, pages 12-21, Edinburgh, Scotland, UK.

Tillmann, C., Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated DP Based Search For Statistical Translation. *Proceedings of the 5th European Conference on Speech Communication and Technology* .

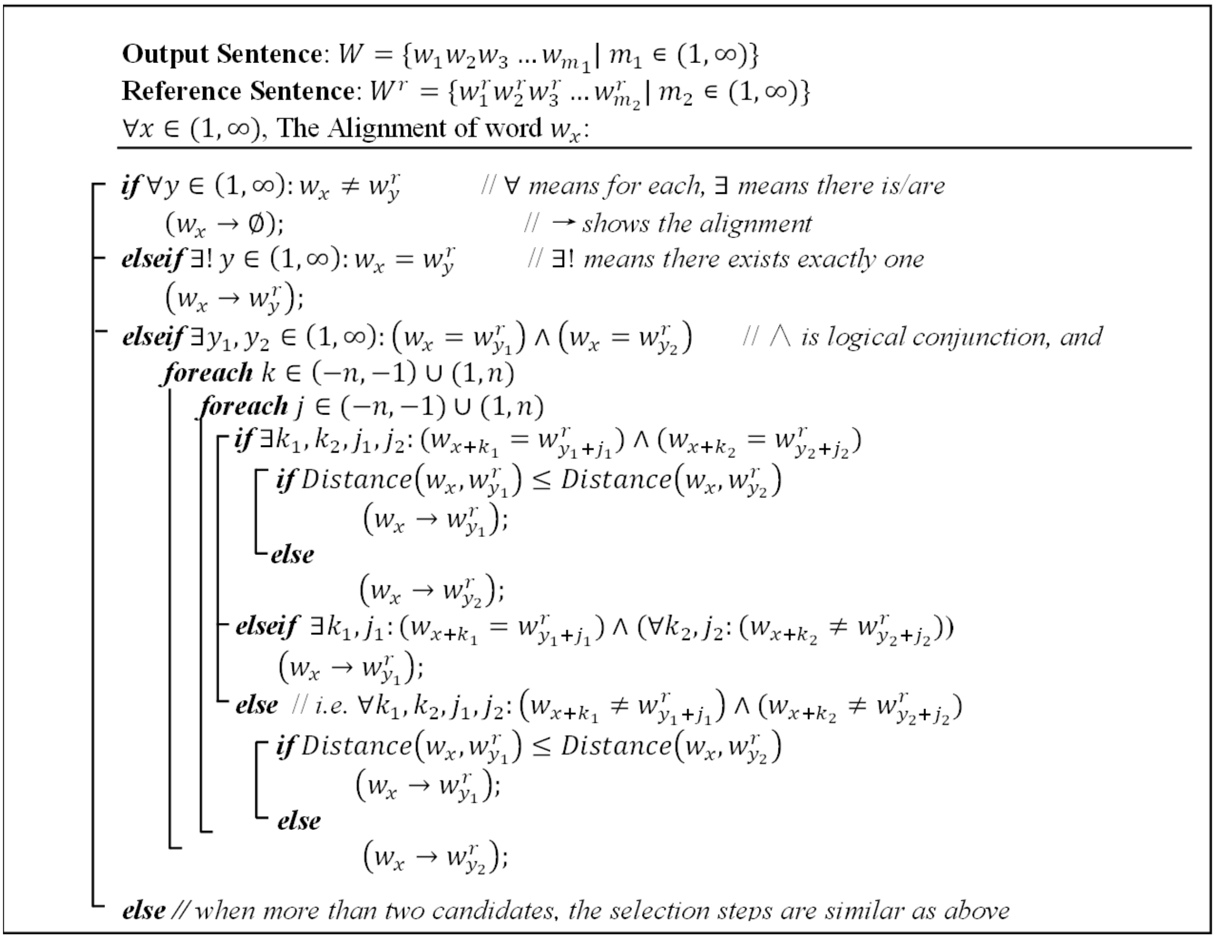
Turian, J. P., Shen, L. and Melanmed, I. D. 2003. Evaluation of machine translation and its evaluation. *Proceedings of MT Summit IX* , pages 386-393, New Orleans, LA, USA.

Weaver, Warren. 1955. Translation. *Machine Translation of Languages: Fourteen Essays*, In William Locke and A. Donald Booth, editors, John Wiley and Sons. New York, pages 15—23.

White, J. S., O'Connell, T. A., and O'Mara, F. E. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. *Proceedings of AMTA*, pp193-205.

Wong, B. T-M and Kit, C. 2008. Word choice and word position for automatic MT evaluation. *Workshop: MetricsMATR of AMTA*, short paper, 3 pages, Waikiki, Hawai'I, USA.

Xiong, D., M. Zhang, H. Li. 2011. A Maximum-Entropy Segmentation Model for Statistical Machine Translation. *IEEE Transactions on Audio, Speech, and Language Processing* , Volume: 19, Issue: 8, 2011 , pp. 2494- 2505.



Output Sentence: $W = \{w_1 w_2 w_3 \dots w_{m_1} | \ m_1 \in (1, \infty)\}$
Reference Sentence: $W^r = \{w_1^r w_2^r w_3^r \dots w_{m_2}^r | \ m_2 \in (1, \infty)\}$
$\forall x \in (1, \infty)$, The Alignment of word w_x:

```
if $\forall y \in (1,\infty): w_x \neq w_y^r$          // $\forall$ means for each, $\exists$ means there is/are
    $(w_x \rightarrow \emptyset)$;                          // → shows the alignment
elseif $\exists!\, y \in (1,\infty): w_x = w_y^r$       // $\exists!$ means there exists exactly one
    $(w_x \rightarrow w_y^r)$;
elseif $\exists y_1, y_2 \in (1,\infty): (w_x = w_{y_1}^r) \wedge (w_x = w_{y_2}^r)$     // $\wedge$ is logical conjunction, and
    foreach $k \in (-n,-1) \cup (1,n)$
        foreach $j \in (-n,-1) \cup (1,n)$
            if $\exists k_1, k_2, j_1, j_2: (w_{x+k_1} = w_{y_1+j_1}^r) \wedge (w_{x+k_2} = w_{y_2+j_2}^r)$
                if $Distance(w_x, w_{y_1}^r) \leq Distance(w_x, w_{y_2}^r)$
                        $(w_x \rightarrow w_{y_1}^r)$;
                else
                        $(w_x \rightarrow w_{y_2}^r)$;
            elseif $\exists k_1, j_1: (w_{x+k_1} = w_{y_1+j_1}^r) \wedge (\forall k_2, j_2: (w_{x+k_2} \neq w_{y_2+j_2}^r))$
                $(w_x \rightarrow w_{y_1}^r)$;
            else // i.e. $\forall k_1, k_2, j_1, j_2: (w_{x+k_1} \neq w_{y_1+j_1}^r) \wedge (w_{x+k_2} \neq w_{y_2+j_2}^r)$
                if $Distance(w_x, w_{y_1}^r) \leq Distance(w_x, w_{y_2}^r)$
                        $(w_x \rightarrow w_{y_1}^r)$;
                else
                        $(w_x \rightarrow w_{y_2}^r)$;
else // when more than two candidates, the selection steps are similar as above
```

Figure 1: N-gram word alignment algorithm

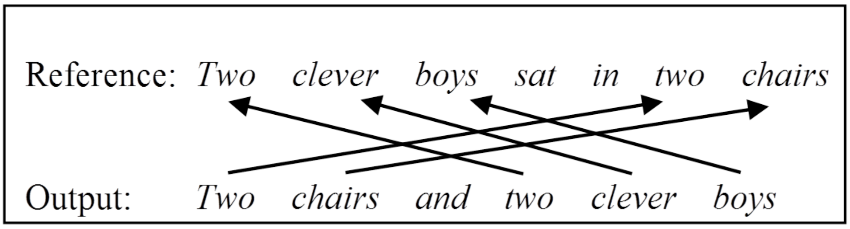
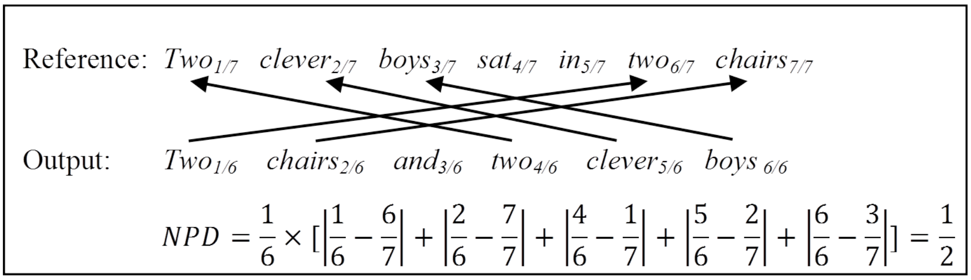

Reference: Two clever boys sat in two chairs
Output: Two chairs and two clever boys

Figure 2: Example of n-gram word alignment


Reference: Two1/7 clever2/7 boys3/7 sat4/7 in5/7 two6/7 chairs7/7
Output: Two1/6 chairs2/6 and3/6 two4/6 clever5/6 boys 6/6

$$NPD = \frac{1}{6} \times [\left|\frac{1}{6} - \frac{6}{7}\right| + \left|\frac{2}{6} - \frac{7}{7}\right| + \left|\frac{4}{6} - \frac{1}{7}\right| + \left|\frac{5}{6} - \frac{2}{7}\right| + \left|\frac{6}{6} - \frac{3}{7}\right|] = \frac{1}{2}$$

Figure 3: Example of NPD calculation