Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 760301, 12 pages
http://dx.doi.org/10.1155/2014/760301

Research Article

Unsupervised Quality Estimation Model for English to German Translation and Its Application in Extensive Supervised Evaluation

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

Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory, Department of Computer and Information Science, University of Macau, Macau

Correspondence should be addressed to Aaron L.-F. Han; hanlifengaaron@gmail.com

Received 30 August 2013; Accepted 2 December 2013; Published 28 April 2014

Academic Editors: J. Shu and F. Yu

Copyright © 2014 Aaron L.-F. Han et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the rapid development of machine translation (MT), the MT evaluation becomes very important to timely tell us whether the MT system makes any progress. The conventional MT evaluation methods tend to calculate the similarity between hypothesis translations offered by automatic translation systems and reference translations offered by professional translators. There are several weaknesses in existing evaluation metrics. Firstly, the designed incomprehensive factors result in language-bias problem, which means they perform well on some special language pairs but weak on other language pairs. Secondly, they tend to use no linguistic features or too many linguistic features, of which no usage of linguistic feature draws a lot of criticism from the linguists and too many linguistic features make the model weak in repeatability. Thirdly, the employed reference translations are very expensive and sometimes not available in the practice. In this paper, the authors propose an unsupervised MT evaluation metric using universal part-of-speech tagset without relying on reference translations. The authors also explore the performances of the designed metric on traditional supervised evaluation tasks. Both the supervised and unsupervised experiments show that the designed methods yield higher correlation scores with human judgments.

1. Introduction

The research about machine translation (MT) can be traced back to fifty years ago [1] and people benefit much from it about the information exchange with the rapid development of the computer technology. Many MT methods and automatic MT systems were proposed in the past years [2–4]. Traditionally, people use the human evaluation approaches for the quality estimation of MT systems, such as the adequacy and fluency criteria. However, the human evaluation is expensive and time consuming. This leads to the appearance of the automatic evaluation metrics, which give quick and cheap evaluation for MT systems. Furthermore, the automatic evaluation metrics can be used to tune the MT systems for better output quality. The commonly used automatic evaluation metrics include BLEU [5], METEOR [6], TER [7], AMBER [8], and so forth. However, most of the automatic MT evaluation metrics are reference aware, which means they tend to employ different approaches to calculate the closeness between the hypothesis translations offered by MT systems and the reference translations provided by professional translators. There are some weaknesses in the conventional reference-aware methods: (1) how many reference translations are enough to avoid the evaluation bias since that reference translations usually cannot cover all the reasonable expressions? (2) The reference translation is also expensive and sometimes not approachable in practice. This paper will propose an automatic evaluation approach for English-to-German translation by calculating the similarity between source and hypothesis translations without using of reference translation. Furthermore, the potential usage of the proposed evaluation algorithms in the traditional reference-aware MT evaluation tasks will also be explored.

2. Traditional MT Evaluations

2.1. BLEU Metric. The commonly used BLEU (bilingual evaluation understudy) metric [5] is designed as automated

substitute to skilled human judges when there is need for quick or frequent MT evaluations:

$$\begin{aligned} \text{BLEU} &= \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log P_n\right), \\ P_n &= \frac{\sum_{C\in\{\text{Candidates}\}} \sum_{n-\text{gram}\in C} \text{Count}_{\text{clip}}\left(n-\text{gram}\right)}{\sum_{C'\in\{\text{Candidates}\}} \sum_{n-\text{gram}'\in C'} \text{Count}\left(n-\text{gram}'\right)}, \end{aligned} \tag{1}$$

where P_n means modified n-gram precision on a multisentence test set, which is for the entire test corpus. It first computes the n-gram matches sentence by sentence and then adds the clipped n-gram counts for all the candidate sentences and divides by the number of candidate n-gram in the test corpus. Consider

$$\text{BP} = \begin{cases} e^{1-r/c} & \text{if } c \le r \\ 1 & \text{if } c > r, \end{cases} \tag{2}$$

where BP is the sentence brevity penalty for short sentences, r is the effective reference sentence length (the closest reference sentence length with the candidate sentence), and c is the length of candidate sentence. Generally, N is selected as 4, and uniform weight w_n is assigned as $1/N$. Thus we could get the deduction as follows:

$$\begin{aligned} \text{BLEU}_{\text{baseline}} &= \text{BP} \times \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log P_n\right) \\ &= \text{BP} \times \exp\left(\frac{1}{N} \log \prod_{n=1}^{N} P_n\right) \\ &= \text{BP} \times \exp\left(\log\left(\prod_{n=1}^{N} P_n\right)^{1/N}\right) \\ &= \text{BP} \times \left(\prod_{n=1}^{N} P_n\right)^{1/N} \\ &= \text{BP} \times \sqrt[N]{\prod_{n=1}^{N} P_n}. \end{aligned} \tag{3}$$

This shows that BLEU reflects the geometric mean of n-gram precision values multiplied by brevity penalty. As a contrast and simplified version, Zhang et al. [9] proposes modified BLEU metric M_BLEU using the arithmetic mean of the n-gram precision:

$$\text{M\_BLEU} = \text{BP} \times \sum_{n=1}^{N} w_n P_n. \tag{4}$$

The weaknesses of BLEU series are that they focus on the usage of incomprehensive factors, precision scores only; they do not use any linguistic features, only utilizing the surface words.

2.2. TER Metric. TER [7] means translation edit rate, which is designed at sentence level to calculate the amount of work needed to correct the hypothesis translation according to the closest reference translation (assuming there are several reference translations):

$$\text{TER} = \frac{\#\text{ of edits}}{\text{average }\#\text{ of reference words}}. \tag{5}$$

The edit categories include the insertion, deletion, substitution of single words, and the shifts of word chunks. TER uses the strict matching of words and word order, for example, miscapitalization is also counted as an edit. The weakness of TER is that it gives an overestimate of the actual translation error rate since that it requests accurate matching between the reference and hypothesis sentence. To address this problem, they proposed the human-targeted TER (HTER) to consider the semantic equivalence, which is achieved by employing human annotators to generate a new targeted reference. However, HTER is very expensive due to that it requires around 3 to 7 minutes per sentence for a human to annotate, which means that it is more like a human judgment metric instead of an automatic one.

2.3. METEOR Metric. METEOR [6] metric conducts a complicated matching, considering stems, synonyms, and paraphrases. Consider

$$\begin{aligned} \text{Score} &= F_{\text{mean}} \times \left(1 - \text{Penalty}\right), \\ \text{Penalty} &= 0.5 \times \left(\frac{\#\text{chunks}}{\#\text{unigrams\_matched}}\right)^3, \\ F_{\text{mean}} &= \frac{10PR}{R + 9P}, \end{aligned} \tag{6}$$

where #chunks means the number of matched chunks between reference and hypothesis sentence, and #unigrams_matched is the number of matched words. It puts more weight on recall (R) compared with precision (P). The matching process involves computationally expensive word alignment due to the external tools for stemming or synonym matches. The advanced version of METEOR is introduced in [10].

2.4. AMBER Metric. AMBER [8] declares a modified version of BLEU. It attaches more kinds of penalty coefficients, combining the n-gram precision and recall with the arithmetic average of F-measure (harmonic mean of precision and recall with the equal weight). It provides eight kinds of preparations on the corpus including whether the words are tokenized or not, extracting the stem, prefix and suffix on the words, and splitting the words into several parts with different ratios. Advanced version of AMBER was introduced in [11]. Other related works about traditional reference-aware MT evaluation metrics can be referred to in the papers [12, 13], our previous works [14, 15], and so forth.

3. Related Works

As mentioned previously, the traditional evaluation metrics tend to estimate quality of the automatic MT output by measuring its closeness with the reference translations. To address this problem, some researchers design the unsupervised MT evaluation approaches without using reference translations, which is also called quality estimation of MT. For example, Han et al. design an unsupervised MT evaluation for French and English translation using their developed universal phrase tagset [16] and explore the performances of machine learning algorithms, for example, conditional random fields, support vector machine and naïve Bayes in the word-level quality estimation task of English to Spanish translation without using golden references [17]. Gamon et al. [18] conduct a research about reference-free MT evaluation approaches also at sentence level, which utilizes the linear and nonlinear combinations of language model and SVM classifier to find the badly translated sentences. Using the regression learning and a set of indicators of fluency and adequacy as pseudoreferences, Albrecht and Hwa [19] present an unsupervised MT evaluation work at sentence level performance. Employing the confidence estimation features and a learning mechanism trained on human annotations, Specia and Giménez [20] develop some quality estimation models, which are biased by difficulty level of the input segment. The issues between the traditional supervised MT evaluations and the latest unsupervised MT evaluations are discussed in the work of [21]. The quality estimation addresses this problem by evaluating the quality of translations as a prediction task and the features are usually extracted from the source sentences and target (translated) sentences. Using the IBM model one and the information of morphemes, lexicon probabilities, part-of-speech, and so forth, Popović et al. [22] also introduces an unsupervised evaluation method and show that the most promising setting comes from the IBM-1 scores calculated on morphemes and POS-4gram. Mehdad et al. [23] use the cross-lingual textual entailment to push semantics into the MT evaluation without using reference translations, which mainly focuses on the adequacy estimation. Avramidis [24] performs an automatic sentence-level ranking of multiple machine translations using the features of verbs, nouns, sentences, subordinate clauses, and punctuation occurrences to derive the adequacy information. Other related works that introduce the unsupervised MT evaluations include [25, 26].

4. Designed Approach

To reduce the expensive reference translations provided by human labor and some external resources such as synonyms, this work employs the universal part-of-speech (POS) tagset containing 12 universal tags proposed by [27]. A part-of-speech is a word class, a lexical class, or a lexical category, which is a linguistic category of words (or lexical items). It is generally defined by the syntactic or morphological behavior of the lexical item in question.

For a simple example, "there is a big bag" and "there is a large bag" could be the same expression, "big" and "large" having the same POS as adjective. To try this potential approach, we conduct the evaluation on the POS of the words from the source language and the target language. The source language is used as pseudoreference. We will also test this method by calculating the correlation coefficient of this approach with the human judgments in the experiment. Petrov et al. [27] describe that the English PennTreebank [28] has 45 tags and German Negra [29] has 54 tags. However, in the mapping table they offer 49 tags for the German Negra Treebank. This paper makes a test on the Berkeley parser [30] for German (trained on the German Negra) parsing a German corpus from the workshop of machine translation (WMT) 2012 [25] and finds that there are indeed other POS tags that are not included in the mapped 49 tags. So, firstly this paper conducts a complementary mapping work for German Negra POS tagset and extends the mapped POS tags to 57 tags.

4.1. Complementary POS Mapping. The parsing test result of WMT 2012 German corpus shows that the omissive German POS tags in the mapping table include "PWAV, PROAV, PIDAT, PWAT, PWS, PRF, \$∗LRB∗, and ∗TN∗," of which "∗TN∗" is a formal style (*N* is replaced in practical parsing results by the integer number such as 1, 2, etc.).

This paper classifies the omissive German POS tags according to the English POS tagset classification since that the English PennTreebank 45 POS tags are completely mapped by the universal POS tagset, as shown in Table 1.

(i) The German POS "PWAV" has a similar function to English POS "WRB" which means wh-adverb labeling the German word such as während (while), wobei (where), wann (when), and so forth. The German POS "PROAV" has a similar function to English POS "RB" which means adverb labeling the German word Dadurch (thereby), dabei (there), and so forth. So this paper classifies "PWAV" and "PROAV" into the ADV (adverb) category in the 12 universal POS tags.

(ii) The German POS "PIDAT" has a similar function to English POS "PDT" which means predeterminer labeling the German word jedem (each), beide (both), meisten (most), and so forth. The German POS "PWAT" has a similar function to English POS "WDT" which means wh-determiner labeling the German word welche (which), welcher (which), and so forth. So "PIDAT" and "PWAT" are classified into the DET (determiner) category in the universal POS tags.

(iii) The German POS "PWS" has a similar function to English POS "WP" which means wh-pronoun labeling the German word was (what), wer (who), and so forth. So "PWS" is classified into the PRON (pronoun) category in the universal POS tags.

(iv) The German POS "PRF" has a similar function to English POS "RP" and "TO", which means particle and to, respectively, labeling the German word sich (itself). So this paper classifies "PRF" into the PRT (a less clear case for particle, possessive, and to) category.

Table 1: Complementary German POS (as bold) mappings for Universal POS tagset.

Language	ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRON	PRT	VERB	X	.
English Penn Treebank [28]	JJ JJR JJS	IN	RB RBR RBS WRB	CC	DT EX PDT WDT	NN NNP NNPS NNS	CD	PRP PRP$ WP WP$	POS RP TO	MD VB VBD VBG VBN VBP VBZ	FW LS SYM UH	# $ ” . : -LRB- -RRB-
German Negra Treebank [29]	ADJA ADJD	APPO APPR APPRART APZR	ADV **PWAV** **PROAV**	KOKOM KON KOUI KOUS	ART **PIDAT** **PWAT**	NE NN NNE	CARD	PDAT PDS PIAT PIS PPER PPOSAT PPOSS PRELAT PRELS **PWS**	PTKA PTKANT PTKNEG PTKVZ PTKZU **PRF**	VAFIN VAIMP VAINF VAPP VMFIN VMINF VMPP VVFIN VVIMP VVINF VVIZU VVPP	FM ITJ TRUNC XY	$($, $. ∗**TN**∗ **$**∗**LRB**∗

Hypothesis Sentence (POS): $P = \{p_1, p_2, \ldots, p_{m1} | m_1 \in (1, \infty)\}$
Source Sentence (POS): $P^s = \{p_1^s, p_2^s, \ldots, p_{m_2}^s | m_2 \in (1, \infty)\}$
$\forall x \in (1, \infty)$, The Alignment of POS p_x in hypothesis:

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

Algorithm 1: The POS alignment algorithm from hypothesis to source sentence based on n-gram.

(v) The German POS "∗TN∗" and "$∗LRB∗" are classified into the punctuation category since that they are only used to label the German punctuations such as dash, bracket, and so forth. After all of the complementary mapping, the universal POS tagset alignment for German Negra Treebank is shown in Table 1 with the boldface POS as the added ones.

4.2. Calculation Algorithms. The designed calculation algorithms of this paper are LEPOR series. First, we introduce the nLEPOR model, n-gram based quality estimation metric for machine translation with augmented factor of enhanced length penalty, precision, position difference penalty, and recall. We will introduce the subfactors in the formula step by step:

$$\begin{aligned} \text{nLEPOR} &= \text{LP} \times \text{NPosPenal} \\ &\quad \times \exp\left(\sum_{n=1}^{N} w_n \log H\left(\alpha R_n, \beta P_n\right)\right) \\ \text{LP} &= \begin{cases} e^{1-s/c} & \text{if } c < s \\ 1 & \text{if } c = s \\ e^{1-c/s} & \text{if } c > s. \end{cases} \end{aligned} \tag{7}$$

In the formula, LP means enhanced sentence length penalty that is designed for both the shorter or longer translated sentences (hypothesis) compared with the source sentence. This approach is different with BLEU metric which assigns penalty for the shorter sentence compared with the human reference translation. Parameters c and s specify the length of candidate sentence (hypothesis) and source sentence, respectively.

The variable NPosPenal means n-gram position difference penalty that is designed for the different order of successful matched POS in source and hypothesis sentence. The position difference factor has been proved to be helpful for the MT evaluation in the research work of [13]. The alignment direction is from hypothesis to the source sentence with the algorithm shown in Algorithm 1. This paper employs the n-gram method into the matching process, which means that the potential POS candidate will be assigned higher priority if it has neighbor matching. The nearest matching will be accepted as a backup choice if there are both neighbor matching or there is no other matched POS around the potential pairs. In Algorithm 1, assuming that p_x represents the current universal POS in hypothesis sentence, p_{x+k} means the kth POS to the previous ($k < 0$) or following ($k > 0$) position. It is the similar approach for the source sentence. Consider

$$\begin{aligned} \text{NPosPenal} &= e^{-\text{NPD}}, \\ \text{NPD} &= \frac{1}{\text{Length}_{\text{hyp}}} \sum_{i=1}^{\text{Length}_{\text{hyp}}} \left|\text{PD}_i\right|, \\ \left|\text{PD}_i\right| &= \left|\text{MatchPosN}_{\text{hyp}} - \text{MatchPosN}_{\text{src}}\right|. \end{aligned} \tag{8}$$

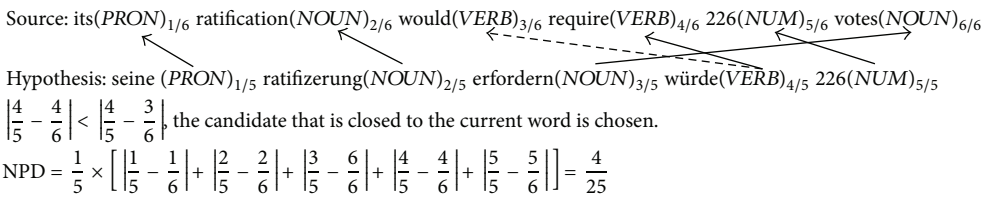


Figure 1: Universal POS alignment and NPD calculation example.

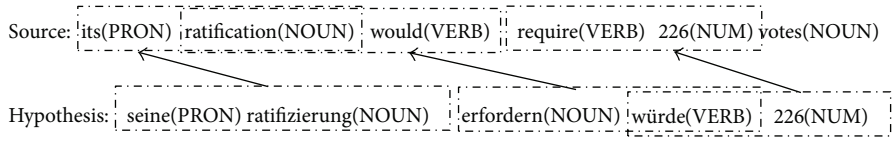


Figure 2: Universal POS chunk matching example for bigram precision and recall.

The parameter $\text{Length}_{\text{hyp}}$ means the length of hypothesis sentence, $\text{MatchPosN}_{\text{hyp}}$, and $\text{MatchPosN}_{\text{src}}$ as the matched POS position number in hypothesis and source sentence, respectively. See Figure 1 as an example.

The parameter w_n is designed to adjust the weights of different n-gram performances such as unigram, bigram, trigram, four-gram, and so forth, which is different with the weight assignment in BLEU where each weight is equal to $1/N$. In our model, higher weight value is designed for the high level n-gram. Consider the following:

$$w_n = \frac{n}{\sum_{i=1}^{N} i} = \frac{n}{1 + 2 + \cdots + N},$$
$$H(\alpha R_n, \beta P_n) = \frac{(\alpha + \beta)}{(\alpha / R_n + \beta / P_n)}. \tag{9}$$

The factor $H(\alpha R_n, \beta P_n)$ is the mathematical harmonic mean of n-gram precision (P_n) and n-gram recall (R_n) in (10), where $\#n\text{gram}_{\text{matched}}$ represents the number of matched n-gram chunks. The n-gram precision (and recall) is calculated on sentence level not corpus-level used in BLEU (P_n). Let us see the example in Figure 1 again for the explanation of bigram precision P_2 and bigram recall R_2. The number of bigram chunks in hypothesis is 4 (PRON NOUN, NOUN NOUN, NOUN VERB, and VERB NUM), the number of bigram chunks in source is 5 (PRON NOUN, NOUN VERB, VERB VERB, VERB NUM, and NUM NOUN), and the number of matched bigram chunk is 3 (PRON NOUN, NOUN VERB, and VERB NUM) as shown in Figure 2. So the value of P_2 and R_2 equals 3/4 and 3/5, respectively.

$$P_n = \frac{\#n\text{gram}_{\text{matched}}}{\#n\text{gram chunks in hypothesis}},$$
$$R_n = \frac{\#n\text{gram}_{\text{matched}}}{\#n\text{gram chunks in source}}. \tag{10}$$

4.3. System-Level Metric. We design two approaches for document-level calculation for the proposed algorithms:

$$\overline{\text{nLEPOR}_A} = \frac{1}{\text{SentNum}} \sum_{i=1}^{\text{SentNum}} \text{NLEPOR}_i,$$
$$\overline{\text{nLEPOR}_B} = \overline{\text{LP}} \times \overline{\text{PosPenalty}} \times \overline{\exp\left(\sum_{n=1}^{N} w_n \log H(\alpha R_n, \beta P_n)\right)}. \tag{11}$$

The document-level $\overline{\text{nLEPOR}_A}$ is calculated as the arithmetic mean value of each sentence-level score in the document. On the other hand, the document-level $\overline{\text{nLEPOR}_B}$ is calculated as the product of three document-level variables, and the document-level variable value is the corresponding arithmetic mean score of each sentence.

5. Usage in Traditional Evaluation

We introduce the potential usage of the proposed evaluation algorithms in traditional supervised (reference-aware) MT evaluation methods. In the unsupervised design, the nLEPOR metric is measured on the source and target POS sequences instead of the surface words. In the exploration of its usage in the traditional supervised MT evaluations, we measure the nLEPOR score on the target translations (system outputs) and reference translations, that is, measuring on the surface words. The pseudoreference (source language) is replaced with real reference here. The performance of simplified variant nLEPOR_A will be tested using the reference translations.

6. Evaluating the Evaluation Method

The conventional method to evaluate the quality of different automatic MT evaluation metrics is to calculate their correlation scores with human judgments. The Spearman

Table 2: Performances on the WMT11 training corpora.

Metrics	UseReference?	rs
AMBER	Yes	0.53
BLEU	Yes	0.44
METEOR	Yes	0.30
$\overline{\text{nLEPOR}_A}$	**No**	**0.63**
$\overline{\text{nLEPOR}_B}$	**No**	0.60

Bold fonts mean the best performance.

Table 3: Performances on the WMT12 testing corpora.

Metrics	UseReference?	rs
AMBER	Yes	0.25
BLEU	Yes	0.22
METEOR	Yes	0.18
$\overline{\text{nLEPOR}_A}$	**No**	**0.34**
$\overline{\text{nLEPOR}_B}$	**No**	0.33

Bold fonts mean the best performance.

rank correlation score (r_s) and Pearson correlation score ρ are commonly used by the annual workshop of statistical machine translation (WMT) of Association for Computational Linguistics (ACL) [25, 31, 32]. Assuming that $\overrightarrow{X} = \{x_1, x_2, \ldots, x_n\}$ and $\overrightarrow{Y} = \{y_1, y_2, \ldots, y_n\}$ are two rank sequences and n is the number of variables, when there are no ties, Spearman rank correlation coefficient is calculated as

$$r_{s_{\theta(XY)}} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}, \tag{12}$$

where d_i is the difference value (D-value) between the two coordinate rank variables $(x_i - y_i)$.

Secondly, the Pearson correlation coefficient information is introduced as below. Given a sample of paired data (X, Y) as (x_i, y_i), $i = 1$ to n, the Pearson correlation coefficient is

$$\rho_{XY} = \frac{\sum_{i=1}^{n}\left(x_i - \mu_x\right)\left(y_i - \mu_y\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \mu_x\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \mu_y\right)^2}}, \tag{13}$$

where μ_x and μ_y specify the arithmetical means of discrete random variable X and Y, respectively.

7. Experiments

7.1. Unsupervised Performances. In the unsupervised MT evaluation, this paper uses the English-to-German machine translation corpora (produced by around twenty English-to-German MT systems) from ACL-SIGMT (http://www.sigmt.org/), which makes the annual workshop corpora public available for further research purpose. Each document contains 3,003 sentences of source English or translated German. To avoid the over fitting problem, the WMT2011 (http://www.statmt.org/wmt11/) corpora are used as training data and WMT2012 (http://www.statmt.org/wmt12/) corpora are used for the testing. This paper conducts the experiments on the simplified version (unigram precision and recall) of the metric.

Table 2 shows the system-level Spearman rank correlation scores of nLEPOR with human judgments trained on the WMT2011 data, as compared to several state-of-the-art reference-aware automatic evaluation metrics including BLEU, METEOR, and AMBER.

In the training period, the parameter values of α (weight on recall) and β (weight on precision) are tuned to 1 and 9, respectively, which is different with the reference-aware metric METEOR (more weight on recall). Bigram is selected for the n-gram universal POS alignment period. The correlation scores in Table 2 show that the proposed evaluation approaches of this paper have achieved higher score (0.63 and 0.60, resp.) in training period than the other evaluation metrics (with the score 0.53, 0.44, and 0.30, resp.) apparently even though the compared metrics are reference aware.

Testing result of the proposed evaluation approaches on the WMT2012 corpora is shown in Table 3 with the same parameter values obtained as in training, also compared with the state-of-the-art evaluation metrics. The correlation scores in Table 3 show the same rank results with Table 2, METEOR achieving the lowest score (0.18), AMBER (0.25) achieving score higher than BLEU (0.22), and NLEPOR family yielding the highest correlation coefficient (0.34 and 0.33, resp.) with human judgments as previous. The test results show the robustness of the proposed evaluation approaches. The experiments also show that the latest proposed metrics (e.g., AMBER) achieve higher correlation score than the earlier ones (e.g., BLEU).

7.2. Supervised Performances. As mentioned previously, to explore the performance of the designed algorithm nLEPOR_A in the traditional reference-aware MT evaluation track, we also use the WMT11 corpora as training data and WMT12 corpora as testing data to avoid the overfitting phenomenon. The number of participated MT systems that offer the output translations is shown in Table 5 for each language pair. In the training period, the tuned values of α and β are 9 and 1, respectively, for all language pairs except for ($\alpha = 1$, $\beta = 9$) for CS-EN. The training results on WMT11 eight corpora including English-to-other (CS-Czech, DE-German, ES-Spanish, FR-French), and other-to-English are shown in Table 4. The aim of the training stage is to achieve higher correlation with human judgments. The experiments on WMT11 corpora show that nLEPOR_A yields the best correlation scores on the language pairs of CS-EN, ES-EN, EN-CS, and EN-ES, which contributes to the highest average score 0.77 on all the eight language pairs. The testing results on WMT12 corpora show that nLEPOR_A yields the highest correlation score with human judgments on CS-EN (0.89) and EN-ES (0.45) language pairs, and the highest average correlation scores on other-to-English (0.85), English-to-other (0.58), and all the eight corpora (0.71) (Table 6).

Table 4: The performances on WMT11 training corpora using Spearman correlation.

System	Correlation score with human judgments										
	Other-to-English					English-to-other					Mean
	CS-EN	DE-EN	ES-EN	FR-EN	Mean	EN-CS	EN-DE	EN-ES	EN-FR	Mean	
nLEPOR$_A$	**0.95**	0.61	**0.96**	0.88	0.85	**0.68**	0.35	**0.89**	0.83	0.69	**0.77**
METEOR	0.91	**0.71**	0.88	**0.93**	**0.86**	0.65	0.30	0.74	0.85	0.64	0.75
BLEU	0.88	0.48	0.90	0.85	0.78	0.65	**0.44**	0.87	**0.86**	**0.71**	0.74
TER	0.83	0.33	0.89	0.77	0.71	0.50	0.12	0.81	0.84	0.57	0.64

Bold fonts mean the best performance.

Table 5: The number of effective MT systems in WMT11 and WMT12.

Year	Number of evaluated MT systems							
	Other-to-English				English-to-other			
	CS-EN	DE-EN	ES-EN	FR-EN	EN-CS	EN-DE	EN-ES	EN-FR
WMT11	8	20	15	18	10	22	15	22
WMT12	6	16	12	15	13	15	11	15

Table 6: The performances on WMT12 testing corpora using Spearman correlation.

System	Correlation score with human judgments										
	Other-to-English					English-to-other					Mean
	CS-EN	DE-EN	ES-EN	FR-EN	Mean	EN-CS	EN-DE	EN-ES	EN-FR	Mean	
nLEPOR$_A$	**0.89**	0.77	0.91	0.81	**0.85**	0.75	0.34	**0.45**	0.77	**0.58**	**0.71**
METEOR	0.66	**0.89**	**0.95**	**0.84**	0.84	0.73	0.18	**0.45**	**0.82**	0.55	0.69
BLEU	**0.89**	0.67	0.87	0.81	0.81	**0.80**	0.22	0.40	0.71	0.53	0.67
TER	**0.89**	0.62	0.92	0.82	0.81	0.69	**0.41**	**0.45**	0.66	0.55	0.68

Bold fonts mean the best performance.

Table 7: The tuned weight values in hLEPOR.

Ratio	Other-to-English					English-to-other				
	CZ-EN	DE-EN	ES-EN	FR-EN	RU-EN	EN-CZ	EN-DE	EN-ES	EN-FR	EN-RU
HPR : LP : NPosPenal (word)	7 : 2 : 1	3 : 2 : 1	7 : 2 : 1	3 : 2 : 1	3 : 2 : 1	7 : 2 : 1	1 : 3 : 7	3 : 2 : 1	3 : 2 : 1	3 : 2 : 1
HPR : LP : NPosPenal (POS)	NA	3 : 2 : 1	NA	3 : 2 : 1	3 : 2 : 1	7 : 2 : 1	7 : 2 : 1	NA	3 : 2 : 1	3 : 2 : 1
$\alpha : \beta$ (word)	1 : 9	9 : 1	1 : 9	9 : 1	1 : 1	9 : 1	9 : 1	9 : 1	9 : 1	1 : 1
$\alpha : \beta$ (POS)	NA	9 : 1	NA	9 : 1	1 : 1	9 : 1	9 : 1	NA	9 : 1	1 : 1
$w_{hw} : w_{hp}$	NA	1 : 9	NA	9 : 1	1 : 1	1 : 9	1 : 9	NA	9 : 1	1 : 1

Table 8: The number of effective MT systems in WMT13.

Year	Number of evaluated MT systems									
	Other-to-English					English-to-other				
	CS-EN	DE-EN	ES-EN	FR-EN	RU-EN	EN-CS	EN-DE	EN-ES	EN-FR	EN-RU
WMT13	12	23	17	19	23	14	21	18	23	19

Table 9: The performances on WMT13 shared tasks using Spearman rank correlation.

System	Correlation score with human judgments											
	Other-to-English						English-to-other					
	CS-EN	DE-EN	ES-EN	FR-EN	RU-EN	Mean	EN-CS	EN-DE	EN-ES	EN-FR	EN-RU	Mean
hLEPOR	0.80	0.93	0.75	0.95	0.79	0.84	0.75	**0.90**	0.84	0.90	**0.85**	**0.85**
nLEPOR$_A$	0.85	0.95	0.83	0.95	0.72	0.86	0.82	**0.90**	**0.85**	**0.92**	0.73	0.84
METEOR	**0.96**	**0.96**	**0.98**	0.98	**0.81**	**0.94**	**0.94**	0.88	0.78	**0.92**	0.57	0.82
BLEU	0.94	0.90	0.88	**0.99**	0.67	0.88	0.90	0.79	0.76	0.90	0.57	0.78
TER	0.80	0.83	0.83	0.95	0.60	0.80	0.86	0.85	0.75	0.91	0.54	0.78

Bold fonts mean the best performance.

Table 10: The performances on WMT13 shared tasks using Pearson correlation.

System	Correlation score with human judgments											
	Other-to-English						English-to-other					
	CS-EN	DE-EN	ES-EN	FR-EN	RU-EN	Mean	EN-CS	EN-DE	EN-ES	EN-FR	EN-RU	Mean
hLEPOR	0.81	0.96	0.90	0.96	0.71	0.87	0.76	**0.94**	**0.91**	0.91	**0.77**	**0.86**
nLEPOR$_A$	0.80	0.94	0.94	0.96	0.69	0.87	**0.82**	0.92	0.90	**0.92**	0.68	0.85
METEOR	**0.99**	**0.96**	**0.97**	**0.98**	**0.84**	**0.95**	**0.82**	0.88	0.88	0.91	0.55	0.81
BLEU	0.89	0.91	0.94	0.94	0.60	0.86	0.80	0.82	0.88	0.90	0.62	0.80
TER	0.77	0.87	0.91	0.93	0.80	0.80	0.70	0.73	0.78	0.91	0.61	0.75

Bold fonts mean the best performance.

7.3. Enhanced Model and the Performances. In the previous sections, we have introduced the n-gram based metric nLEPOR and its performances in both supervised and unsupervised cases. In this section we will introduce an enhanced version of the proposed metric, which is called as hLEPOR (harmonic mean of enhanced length penalty, precision, n-gram position difference penalty, and recall). There are two contributions of this enhanced model. Firstly, it assigns different weights to three subfactors, which are tunable according to different language pairs. This property can help to address the language-bias problem existing in many current automatic evaluation metrics. Secondly, it is designed to combine the performances on words and POS together, and the final score is the combination of them:

$$\text{hLEPOR} = \frac{w_{\text{LP}} + w_{\text{NPosPenal}} + w_{\text{HPR}}}{w_{\text{LP}}/\text{LP} + w_{\text{NPosPenal}}/\text{NPosPenal} + w_{\text{HPR}}/\text{HPR}}, \tag{14}$$

where HPR is the harmonic mean of precision and recall as mentioned above $H(\alpha R_n, \beta P_n)$. Consider

$$\text{hLEPOR}_{\text{final}} = \frac{1}{w_{\text{hw}} + w_{\text{hp}}} \times \left(w_{\text{hw}}\text{hLEPOR}_{\text{word}} + w_{\text{hp}}\text{hLEPOR}_{\text{POS}}\right). \tag{15}$$

Firstly, we calculate the hLEPOR score on surface words $\text{hLEPOR}_{\text{word}}$, that is, the closeness of the hypothesis translation and the reference translation. Then we calculate the hLEPOR score on the extracted POS sequences $\text{hLEPOR}_{\text{POS}}$, that is, the closeness of the corresponding POS tags between hypothesis sentence and reference sentence. The final score $\text{hLEPOR}_{\text{final}}$ is the combination of the two subscores $\text{hLEPOR}_{\text{word}}$ and $\text{hLEPOR}_{\text{POS}}$.

We introduce the performances of nLEPOR and hLEPOR in the WMT13 (http://www.statmt.org/wmt13/) shared evaluation tasks below. Both of the two metrics are trained on WMT11 corpora. The tuned parameters of hLEPOR are shown in Table 7, using default values for EN-RU and RU-EN. The number of MT systems for each language pair in WMT13 is shown in Table 8. There is a new language Russian in WMT13, which leads to the increasing of total number of corpora into ten. Due to the fact that there is no Russian language in the past WMT shared tasks, for EN-RU and RU-EN corpora, we assign the default values $\alpha = 1$ and $\beta = 1$ in nLEPOR.

In the WMT13 shared tasks, both Pearson correlation coefficient and the Spearman rank correlation coefficient are used as the evaluation criteria. So we list the official results in Tables 9 and 10, respectively, by using Spearman rank correlation and Pearson correlation criteria.

Using the Spearman rank correlation coefficient, the experiment results on WMT13 in Table 9 show that hLEPOR yields the highest correlation scores on EN-DE (0.90), EN-RU (0.85), and the highest average correlation score (0.85) on five English-to-other corpora; nLEPOR$_A$ yields the highest correlation scores on EN-DE (0.90), EN-ES (0.85), and EN-FR (0.92) and the second highest average correlation score (0.84) on five English-to-other corpora.

Using the Pearson correlation coefficient, the experiment results on WMT13 in Table 10 show that hLEPOR yields the highest correlation scores on EN-DE (0.94), EN-ES (0.91), and EN-RU (0.77) and the highest average correlation score (0.86) on five English-to-other corpora; nLEPOR$_A$ yields the highest correlation scores on EN-ES (0.82) and EN-FR (0.92) and the second highest average correlation score (0.85) on five English-to-other corpora.

On the other hand, METEOR yields the best performances on other-to-English translation evaluation direction. This is due to the fact that METEOR employs many external materials including stemming, synonyms vocabulary, paraphrasing resources, and so forth. However, to make the evaluation model concise, our method nLEPOR only uses the surface words and hLEPOR only uses the combination of surface words and POS sequences. This shows the advantages of our designed methods, that is, concise linguistic feature.

As mentioned previously, the language-bias problem is presented in the evaluation results of BLEU metric. BLEU yields the highest Spearman rank correlation score 0.99 on FR-EN; however it never achieves the highest average correlation score on any translation direction, due to the very low correlation scores on RU-EN, EN-DE, EN-ES, and EN-RU corpora. Our metrics LEPOR series address the language-bias problem by using augmented factors and tunable parameters.

The evaluation results in Tables 9 and 10 show that the evaluation on the language pairs with English as the source language (English-to-other) is the main challenge at

Table 11

POS	Frequency
$∗LRB∗	213
$,	4590
$.	3270
∗**T1**∗	68
∗**T2**∗	11
∗**T3**∗	3
∗**T4**∗	1
ADJA	4337
ADJD	1943
ADV	2900
APPO	14
APPR	5325
APPRART	1187
APZR	29
ART	7748
CARD	1135
FM	172
ITJ	1
KOKOM	166
KON	1943
KOUI	136
KOUS	929
NE	4102
NN	15211
PDAT	340
PDS	292
PIAT	252
PIDAT	292
PIS	824
PPER	1827
PPOSAT	703
PRELAT	38
PRELS	810
PRF	471
PROAV	280
PTKA	37
PTKANT	5
PTKNEG	393
PTKVZ	433
PTKZU	476
PWAT	14
PWAV	240
PWS	86
TRUNC	18
VAFIN	2197
VAINF	202
VAPP	84
VMFIN	417
VMINF	4
VVFIN	3818
VVIMP	18
VVINF	1051
VVIZU	173
VVPP	1408
XY	14

Table 11: Continued.

system-level performance. Fortunately, the designed evaluation methods of this paper have made some contributions on this translation evaluation direction.

8. Discussion

The Spearman rank correlation coefficient is commonly used in the WMT shared tasks as the special case of Pearson correlation coefficient applied to ranks. However, there is some information that cannot be reflected by using Spearman rank correlation coefficient instead of Pearson correlation coefficient. For example, let us assume there are three automatic MT systems $\overrightarrow{M} = \{M_1, M_2, M_3\}$ and two automatic MT evaluation systems MTE_A and MTE_B. The evaluation scores of the two evaluation systems on the three MT systems are $\overrightarrow{\mathrm{MTE}_A} = \{0.90, 0.35, 0.40\}$ and $\overrightarrow{\mathrm{MTE}_B} = \{0.46, 0.35, 0.42\}$, respectively. Using the Spearman rank correlation coefficient, the two vectors will be first converted into $\widetilde{\overrightarrow{\mathrm{MTE}_A}} = \{1, 3, 2\}$ and $\widetilde{\overrightarrow{\mathrm{MTE}_B}} = \{1, 3, 2\}$, respectively, then their correlation with human rank results (human judgments) will be measured. Thus, the two evaluation metrics will yield the same Spearman rank correlation score with human judgments no matter what the manual evaluation results will be. However, the evaluation systems MTE_A and MTE_B actually tell different results on the quality of the three MT systems. The evaluation system MTE_A gives very high score 0.90 on M_1 system and similar low scores 0.35 and 0.40, respectively, on M_2 and M_3 systems. On the other hand, the evaluation system MTE_B yields similar scores 0.46, 0.35, and 0.42, respectively, on the three MT systems, which means that all the three automatic MT systems have low translation quality. Using the Spearman rank correlation coefficient, the above important information is lost.

On the other hand, the Pearson correlation coefficient uses the absolute scores yielded by the automatic MT evaluation systems as shown in (13), without the preconverting into rank values.

9. Conclusions and Future Works

To avoid the usage of expensive reference translations, this paper designs a novel unsupervised MT evaluation model for English to German translation by employing augmented factors and universal POS tagset. Furthermore, the proposed unsupervised model yields higher correlation score with

human judgments as compared to the reference-aware metrics METEOR, BLEU, and AMBER.

The application of the designed algorithms to the traditional supervised evaluation tasks is also explored. To address the language-bias problem in most of the existing metrics, tunable parameters are assigned to different subfactors. The experiment on WMT11 and WMT12 corpora shows that our designed algorithms yields the highest average correlation score on eight language pairs as compared to the state-of-the-art reference-aware metrics METEOR, BLEU, and TER.

On the other hand, to address the linguistic-extreme problem (no linguistic information or too many linguistic features), our method utilizes the optimized linguistic feature POS sequence, in addition to the surface words, to make the model concise and easy to repeat.

Last, this paper also makes a contribution to the complementary POS tagset mapping between German and English in the light of 12 universal tags.

The developed algorithms in this paper are freely available for research purpose (https://github.com/aaronlifenghan/aaron-project-lepor). In the future works, to test the robustness of the designed algorithms and models, we will seek more language pairs, such as the Asian languages Chinese, Korean, and Japanese, to conduct the experiments, in addition to the official European languages offered by SIGMT association. Secondly, the experiments using multireferences will be considered. Thirdly, how to handle the MT evaluation from the aspect of semantic similarity will be further explored.

Appendix

This appendix offers the POS tags and their occurrences number in one of WMT2012 German documents containing 3,003 sentences and parsed by Berkeley parser for German language, that is, trained on German Negra Treebank.

Number of different POS tags in the document is 55. Detailed POS labels (boldface POSs are not covered in the original mapping) and their frequencies are given in Table 11.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors are grateful to the Science and Technology Development Fund of Macau and the Research Committee of the University of Macaufor the funding support for their research, under the Reference nos.MYRG076(Y1-L2)-FST13-WF, and MYRG070(Y1-L2)-FST12-CS.

References

[1] W. Weaver, "Translation," in *Machine Translation of Languages: Fourteen Essays*, W. Locke and A. Donald Booth, Eds., pp. 15–23, 1955.

[2] F. J. Och, "Minimum error rate training for statistical machine translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 160–167, ACL Press, July 2003.

[3] J. B. Marino, R. E. Banchs, J. M. Crego et al., "N-gram-based machine translation," *Computational Linguistics*, vol. 32, no. 4, pp. 527–549, 2006.

[4] G. Lembersky, N. Ordan, and S. Wintner, "Language models for machine translation-original vs. translated texts," *Journal of Computational Linguistics*, vol. 38, no. 4, 2012.

[5] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, ACL Press, 2002.

[6] S. Banerjee and A. Lavie, "METE-OR: an Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the 43th Annual Meeting of the Association of Computational Linguistics*, pp. 65–72, ACL Press, 2005.

[7] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA '06)*, pp. 223–231, AMTA Press, August 2006.

[8] B. Chen and R. Kuhn, "Amber: a modi-fied bleu, enhanced ranking metric," in *Proceedings of the 6th Workshop on Statistical Machine Translation of the Association for Computational Linguistics*, pp. 71–77, ACL Press, 2011.

[9] Y. Zhang, S. Vogel, and A. Waibel, "Interpreting BLEU/NIST scores: how much improvement do we need to have a better system?" in *Proceedings of the Language Resources and Evaluation (LREC '04)*, ELRA Press, Lisbon, Portugal, 2004.

[10] M. Denkowski and A. Lavie, "Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems," in *Proceedings of the 6th Workshop on Statistical Machine translation of the Association for Computational Linguistics (ACL-WMT '11)*, pp. 85–91, ACL Press, 2011.

[11] B. Chen, R. Kuhn, and G. Foster, "Improving AMBER, an MT evaluation metric," in *Proceedings of the 7th Workshop on Statistical Machine translation of the Association for Computational Linguistics (ACL-WMT '12)*, pp. 59–63, ACL Press, 2012.

[12] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT '02)*, pp. 138–145, Morgan Kaufmann, 2002.

[13] B. Wong and C. Kit, "ATEC: automatic evaluation of machine translation via word choice and word order," *Machine Translation*, vol. 23, no. 2-3, pp. 141–155, 2009.

[14] A. L. F. Han, D. F. Wong, and L. S. Chao, "LEPOR: a robust evaluation metric for machine translation with augmented factors," in *Proceedings of the 24th International Conference on Computational Linguistics*, pp. 441–450, International Committee of Computational Linguistics Press, December 2012.

[15] A. L. F. Han, D. F. Wong, L. S. Chao et al., "Language-independent model for machine translation evaluation with reinforced factors," in *Proceedings of the 14th International Conference of Machine Translation Summit*, pp. 215–222, International Association for Machine Translation Press, September 2013.

[16] A. L. F. Han, D. F. Wong, L. S. Chao, L. He, S. Li, and L. Zhu, "Phrase tagset mapping for French and English treebanks and

its application in machine translation evaluation," in *Language Processing and Knowledge in the Web*, vol. 8105 of *Lecture Notes in Computer Science*, pp. 119–131, 2013.

[17] A. L. F. Han, Y. Lu, D. F. Wong, L. S. Chao, L. He, and J. Xing, "Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling," in *Proceedings of the 8th Workshop on Statistical Machine Translation (ACL-WMT '13)*, pp. 365–372, Association for Computational Linguistics Press, August 2013.

[18] M. Gamon, A. Aue, and M. Smets, "Sentence-level MT evaluation without reference translations: beyond language modeling," in *Proceedings of the 10th Annual Conference on European Association for Machine Translation (EAMT '05)*, pp. 103–111, EAMT Press, May 2005.

[19] J. S. Albrecht and R. Hwa, "Regression for sentence-level MT evaluation with pseudo references," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL '07)*, pp. 296–303, ACL Press, June 2007.

[20] L. Specia and J. Giménez, "Combining confidence estimation and reference-based metrics for segment-level MT evaluation," in *Proceedings of the 9th Biennial Conference of the Association for Machine Translation in the Americas (AMTA '10)*, AMTA Press, November 2010.

[21] L. Specia, D. Raj, and M. Turchi, "Machine translation evaluation versus quality estimation," *Machine Translation*, vol. 24, no. 1, pp. 39–50, 2010.

[22] M. Popović, D. Vilar, E. Avramidis, and A. Burchardt, "Evaluation without references: IBM1 scores as evaluation metrics," in *Proceedings of the 6th Workshop on Statistical Machine Translation, Association for Computational Linguistics (ACL-WMT '11)*, pp. 99–103, ACL Press, 2011.

[23] Y. Mehdad, M. Negri, and M. Federico, "Match without a referee: evaluating MT adequacy without reference translations," in *Proceedings of the 7th Workshop on Statistical Machine Translation*, ACL Press, 2012.

[24] E. Avramidis, "Comparative quality estimation: automatic sentence-level ranking of multiple machine translation outputs," in *Proceedings of 24th International Conference on Computational Linguistics (COLING '12)*, pp. 115–132, ACL Press, 2012.

[25] C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia, "Findings of the 2012 workshop on statistical machine translation," in *Proceedings of the 7th Workshop on Statistical Machine Translation*, pp. 10–51, ACL Press, 2012.

[26] M. Felice and L. Specia, "Linguistic features for quality estimation," in *Proceedings of the 7th Workshop on Statistical Machine Translation*, pp. 96–103, ACL Press, 2012.

[27] S. Petrov, D. Das, and R. McDonald, "A universal part-of-speech tagset," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC '12)*, European Language Resources Association (ELRA) Press, 2012.

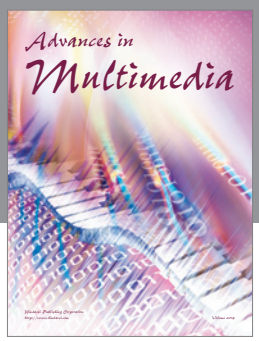
[28] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: the Penn Treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

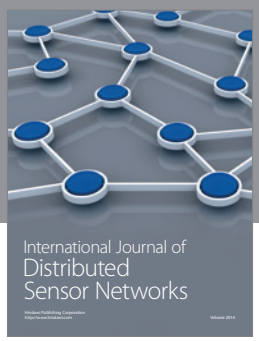
[29] W. Skut, B. Krenn, T. Brants, and H. Uszkoreit, "An annotation scheme for free word order languages," in *Proceedings of the Applied Natural Language Processing*, ACL Press, 1997.

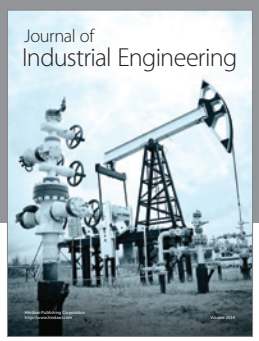
[30] S. Petrov, L. Barrett, R. Thibaux, and D. Klein, "Learning accurate, compact, and interpretable tree annotation," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL '06)*, pp. 433–440, ACL Press, July 2006.

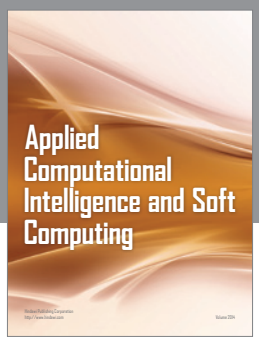
[31] M. Macháček and O. Bojar, "Results of the WMT13 metrics shared task," in *Proceedings of the 8th Workshop on Statistical Machine Translation*, pp. 45–51, ACL Press, 2013.

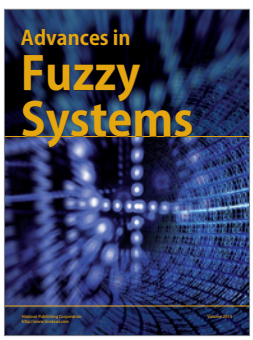
[32] C. Callison-Burch, P. Koehn, C. Monz, and O. F. Zaidan, "Findings of the 2011 workshop on statistical machine translation," in *Proceedings of the 6th Workshop on Statistical Machine Translation of the Association for Computational Linguistics (ACL-WMT '11)*, pp. 22–64, ACL Press, 2011.











Hindawi

Submit your manuscripts at
http://www.hindawi.com