Towards Annotation-Free Evaluation of Cross-Lingual Image Captioning

Aozhu Chen, Xinyi Huang, Hailan Lin, Xirong Li*
MOE Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China
Beijing, China

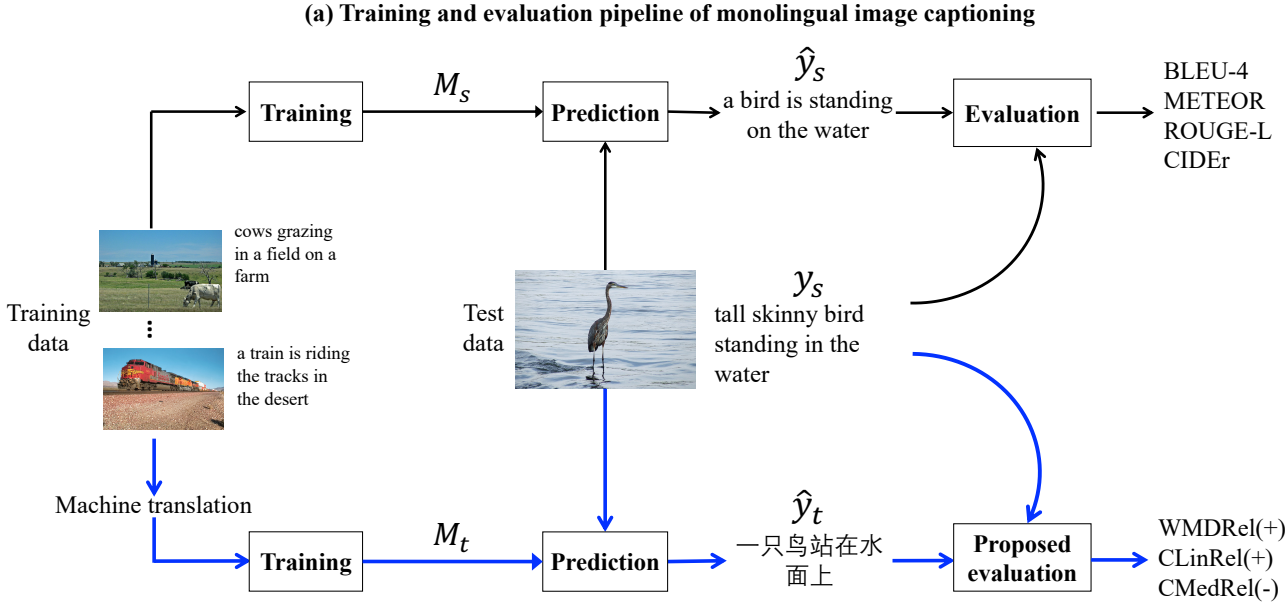


Figure 1: Conceptual diagram of training and evaluation pipelines of (a) *monolingual* image captioning, where training and test data are described by the same language (English) and (b) *cross-lingual* image captioning, where training data is described by a source language (English) while the test data is to be annotated by sentences $\hat{y}_t$ in a distinct target language (Chinese). This paper makes a novel attempt to evaluate the effectiveness of a cross-lingual image captioning model M_t with no need of any reference sentence in the target language. The symbol (+) means the computation of a proposed metric (WMDRel or CLinRel) requires reference y_s in the source language, while (-) means reference-free.

ABSTRACT

Cross-lingual image captioning, with its ability to caption an unlabeled image in a target language other than English, is an emerging topic in the multimedia field. In order to save the precious human resource from re-writing reference sentences per target language, in this paper we make a brave attempt towards annotation-free evaluation of cross-lingual image captioning. Depending on whether we assume the availability of English references, two scenarios are investigated. For the first scenario with the references available, we propose two metrics, i.e., WMDRel and CLinRel. WMDRel measures the semantic relevance between a model-generated caption and machine translation of an English reference using their Word Mover's Distance. By projecting both captions into a deep visual feature space, CLinRel is a visual-oriented cross-lingual relevance measure. As for the second scenario, which has zero reference and is thus more challenging, we propose CMedRel to compute a cross-media relevance between the generated caption and the image content, in the same visual feature space as used by CLinRel. We have conducted a number of experiments to evaluate the effectiveness of the three proposed metrics. The combination of WMDRel, CLinRel and CMedRel has a Spearman's rank correlation of 0.952 with the sum of BLEU-4, METEOR, ROUGE-L and CIDEr, four standard metrics

*This work was supported by NSFC (No. 61672523), BJNSF (No. 4202033), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 18XNLG19). Corresponding author: Xirong Li (xirong@ruc.edu.cn).

MMAsia '20, March 7–9, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8308-0/21/03…$15.00
https://doi.org/10.1145/3444685.3446322

computed using references in the target language. CMedRel alone has a Spearman's rank correlation of 0.786 with the standard metrics. The promising results show high potential of the new metrics for evaluation with no need of references in the target language.

KEYWORDS

Cross-lingual image captioning, evaluation metrics

ACM Reference Format:
Aozhu Chen, Xinyi Huang, Hailan Lin, Xirong Li. 2021. Towards Annotation-Free Evaluation of Cross-Lingual Image Captioning. In *ACM Multimedia Asia (MMAsia '20), March 7–9, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3444685.3446322

1 INTRODUCTION

Image captioning, which aims to automatically describe the pictorial content of an unlabeled image with a sentence, is being actively studied [2, 9, 24]. As its subtopic, cross-lingual image captioning, with the ability to caption a given image in a target language other than English, is attracting an increasing amount of attention in both multimedia and computer vision fields [6, 7, 14, 15, 22, 26].

Previous works on the topic emphasize novel algorithms that effectively learn image captioning models for the target language from existing English datasets such as Flickr8k [8], Flickr30k [28] and MS-COCO [3]. In [15], for instance, Li *et al.* use machine translation to automatically translate English captions of Flickr8k into Chinese and subsequently train a Show-Tell model [24] on the translated dataset. Observing the phenomenon that machine-translated sentences can be unreadable, Lan *et al.* [14] introduce fluency-guided learning, wherein the importance of a training sentence is weighed by its fluency score estimated by a deep language model. Song *et al.* [22] improve [14] by introducing self-supervised reward with respect to both fluency and visual relevance. Although such a training process requires only a small (or even zero) amount of data in the target language, a large-scale evaluation of the resultant models typically needs thousands of test images associated with manually written captions, known as *references*, in the same language. Even assisted by an interactive annotation system [10], months of human labor are required to re-annotate a medium-sized testset per target language.

In this paper we contribute to cross-lingual image captioning with a novel approach to its evaluation. More specifically, we make a brave attempt to remove the need of references in the target languages. We propose three metrics that allow us to differentiate between good-performing and bad-performing models, when a test image is provided with just one reference in English. Such a prerequisite is valid, as the previous works on cross-lingual image captioning are conducted mostly on established English datasets. Our major conclusions are twofold:

- To the best of our knowledge, this is the first work on evaluating image captioning models in a cross-lingual setting, with no need of any reference in the target language. To that end, we propose three metrics, *i.e.,* WMDRel, CLinRel and CMedRel, that assess the semantic relevance of auto-generated captions with respect to the image content in varied manners.
- We have conducted a number of experiments to evaluate the effectiveness of the three proposed metrics. Given the varied combinations of image captioning networks (Show-Tell [24], Up-Down [2] and AoANet [9]) and training datasets (COCO-CN [16] and VATEX [25]) we build a set of eight Chinese models to be ranked. The combination of WMDRel, CLinRel and CMedRel has Spearman's rank correlation of 0.952 with the sum of the four standard metrics, *i.e.,* BLEU-4, METEOR, ROUGE-L and CIDEr. When no reference in the source language is given, CMedRel alone has Spearman correlation of 0.881 with CIDEr.

2 RELATED WORK

We shall clarify that this paper is not about building a better cross-lingual image captioning model. Rather, we are interested in novel metrics that can be computed without the need of reference sentences in a target language.

According to the evaluation protocol used in [14] and its follow-ups, human resources regarding the evaluation of cross-lingual image captioning are spent on two parts. The first part is to manually write references in the target language so that stanard metrics such as BLEU-4 [13], METEOR [4], ROUGE-L [18] and CIDEr [23] can be computed by performing word-level or phrase-level comparison between the auto-generated captions and the references. The second part is to manually assess subjective attributes of sentences such as their readability and fluency. Our proposed approach is to remove the first part so that the relatively limited human resources can be fully spent on the second part. The starting point of our work differs fundamentally from previous efforts on devising better automated metrics [1, 12], as they still assume the availability of references in the target language.

3 PROPOSED APPROACH

3.1 Problem Formalization

A cross-lingual image captioning model M_t in its training stage shall learn from training data described in a source language. While in the inference stage, the model generates for a novel image x a descriptive sentence in a target language, denoted as $\hat{y}_t$:

$$\hat{y}_t \leftarrow M_t(x). \tag{1}$$

When coming to the evaluation stage, the current setting of cross-lingual image captioning [14, 16, 21] assumes the availability of at least one ground-truth sentence in the target language, denoted as y_t, w.r.t the image. Similarly, we use y_s to denote a ground-truth sentence in the source language. Accordingly, the quality of $\hat{y}_t$ is measured based on its word- or phrase- level matching with y_t. Such a matching is typically implemented as $\phi(\hat{y}_t, y_t)$, with $\phi \in$ {BLEU-4, METEOR, ROUGE-L, CIDEr}. Given two distinct models $M_{t,1}$ and $M_{t,2}$, $\phi(M_{t,1}(x), y_t) > \phi(M_{t,2}(x), y_t)$ means the former is better and vice versa. Our goal is to remove the need of y_t.

Depending on whether y_s is available, we consider the following two scenarios:

- Scenario-I: Evaluating M_t on an established dataset with y_s available. This scenario applies to the majority of the works on cross-lingual image captioning, as they evaluate on (a subset) of MS-COCO.

- Scenario-II: Evaluating M_t on a novel and fully unlabeled dataset. This scenario is more practical yet much more challenging.

For Scenario-I, a cross-lingual version of ϕ, indicated by $\phi_{CLin}(\hat{y}_t, y_s)$ is required to measure to what extent $M_t(x)$ matches with y_s. As for Scenario-II, a cross-media version of ϕ, denoted as $\phi_{CMed}(\hat{y}_t, x)$, is needed to measure how $M_t(x)$ matches with the visual content. Note that when comparing distinct models, their rank matters. Hence, the purpose of ϕ_{CLin} and ϕ_{CMed} is to approximate the model rank determined by ϕ. To that end, we develop three metrics, *i.e.,* WMD Relevance (WMDRel) and Cross-Lingual Relevance (CLinRel) to realize ϕ_{CLin}, and Cross-Media Relevance (CMedRel) for ϕ_{CMed}. The three metrics are illustrated in Fig. 2 and depicted as follows.

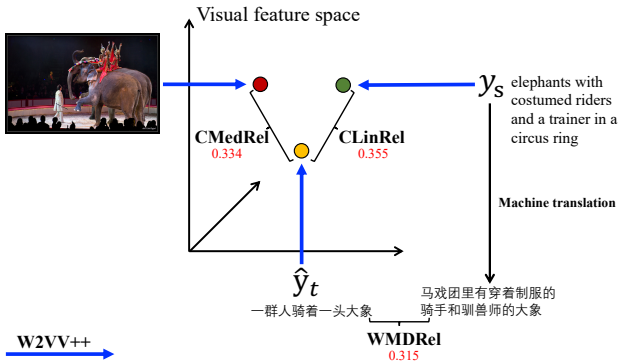


Figure 2: Conceptual illustration of the three proposed metrics. Given a caption $\hat{y}_t$ generated by a cross-lingual image captioning model, we propose WMDRel and CLinRel to measure the semantic relevance between $\hat{y}_t$ and y_s, the reference in a source language (English here), and CMedRel to measure the semantic relevance between $\hat{y}_t$ and the visual content. Different from previous works, no reference caption in the target language (Chinese here) is needed.

3.2 Three Proposed Metrics

3.2.1 WMDRel: Word Mover's Distance based Relevance. We repurpose the Word Mover's Distance (WMD), originally proposed by Kilickaya *et al.* for measuring document similarity [12], in the new context of cross-lingual image captioning evaluation. In order to deal with synonyms and semantically close words that cannot be modeled by bag-of-words based matching, WMD formulates the matching problem between two documents as the classical Earth Mover process, with the goal of moving each word in a document to words in another document. The moving cost between two words is defined as the Euclidean distance between their word2vec features. Accordingly, WMD between two sentences is defined as the minimum cumulative cost of moving all words in one sentence to successfully match with the other sentence.

Note that WMD is monolingual. Therefore, we have y_s automatically translated to the target language (which is Chinese in this study) by machine translation. We use $MT(y_s)$ to indicate the translated reference, and $wmd(\hat{y}_t, MT(y_s))$ as the WMD between $\hat{y}_t$ and $MT(y_s)$. Accordingly, we compute WMDRel as the normalized inverse of $wmd(\hat{y}_t, MT(y_s))$:

$$WMDRel(\hat{y}_t, y_s) = 1 - \frac{wmd(\hat{y}_t, MT(y_s))}{z}, \tag{2}$$

where z is a normalization factor to ensure a score between 0 to 1. A Chinese word2vec model[1], pre-trained on 120G text corpus with 6.1 million tokens, is used.

3.2.2 CLinRel: Cross-Lingual Relevance in Visual Feature Space. It is worth noting that errors in machine translation remain inevitable. As a consequence, $MT(y_s)$ does not fully reflect the semantic meaning of y_s. We therefore look for alternatives that can measure the semantic relevance between $\hat{y}_t$ and y_s with no need of machine translation. Since a visual feature space is naturally cross-lingual, we consider project both $\hat{y}_t$ and y_s into such a feature space and consequently compute their relevance in the common space.

In the context of image/video caption retrieval, Dong *et al.* propose to project a given sentence into a visual feature space by a deep learning model called Word2VisualVec (W2VV) [5]. In particular, the given sentence is first vectorized by three sentence encoders in parallel, *i.e.,* bag-of-words, word2vec and GRU. The output of the encoders is concatenated into a long vector, which is then embedded into the visual feature space by an MLP network. In this work, we adopt W2VV++ [17], a super version of W2VV. We train an English version of W2VV++ and a Chinese version, which are used to project y_s and $\hat{y}_t$ into the visual feature space, respectively. Given $v(y_s)$ and $v(\hat{y}_t)$ as their corresponding vectors, we define CLinRel as their cosine similarity, *i.e.,*

$$CLinRel(\hat{y}_t, y_s) = \frac{v^T(y_s) \cdot v(\hat{y}_t)}{||v^T(y_s)|| \cdot ||v(\hat{y}_t)||}. \tag{3}$$

We instantiate the visual feature space by extracting 2,048-dimensional CNN features using a pre-trained ResNeXt-101 [20], unless stated otherwise.

3.2.3 CMedRel: Cross-Media Relevance. To deal with Scenario-II where y_s is unavailable, we now introduce CMedRel, which assesses $\hat{y}_t$ with respect to the visual content. We compute such cross-modal relevance as the cosine similarity between $v(\hat{y}_t)$ and $v(x)$:

$$CMedRel(\hat{y}_t, x) = \frac{v^T(\hat{y}_t) \cdot v(x)}{||v(\hat{y}_t)|| \cdot ||v(x)||}. \tag{4}$$

4 EVALUATION

4.1 Experimental Setup

We verify the effectiveness of the proposed metrics by evaluating their consistency with the standard metrics, *i.e.,* BLEU-4, METEOR, ROUGE-L, CIDEr and their combination, which are computed based on references in the target language. Given a set of cross-lingual image captioning models, the consistency between two metrics A and B is measured in terms of the Spearman's rank correlation coefficient between model ranks given by A and B. Spearman correlation of +1 means the two metrics are fully consistent.

[1] https://weibo.com/p/23041816d74e01f0102x77v

In what follows, we describe how to build a set of models followed by implementation details.

4.1.1 Model Pool Construction. An image captioning model is determined by two major factors, *i.e.,* network architecture and training data. By trying varied combinations of the two factors, we construct a pool of eight distinct models as follows.

Choices of Training Data. We use the following bilingual (English-Chinese) datasets, wherein Chinese captions are obtained either by machine translation of the original English captions or by manual annotation:

- **COCO-CN** [16]: A public dataset extending MS-COCO with manually written Chinese sentences. It contains 20,342 images annotated with 27,218 Chinese sentences. We use its development set *COCO-CN-dev* as training data.
- **COCO-MT**: Also provided by [16], using the Baidu translation API to automatically translate the original English sentences of MS-COCO to Chinese. COCO-MT contains 123,286 images and 608,873 machine-translated Chinese sentences.
- **VATEX** [25]. A subset of the kinetics-600 [11] short-video collection, showing 600 kinds of human activities. Each video is associated with 10 English sentences and 10 Chinese sentences obtained by crowd sourcing. Following the notation of [16], we term the dataset with only Chinese annotations as *VATEX-CN*. We also construct a machine-translated counterpart, which we term *VATEX-MT*.

We use each of the four datasets, *i.e.,* COCO-CN-dev, COCO-MT, VATEX-CN and VATEX-MT, as training data. Basic statistics of the datasets and their usage in our experiments are given in Table 1.

Table 1: Datasets used in our experiments. A dataset postfixed with "-MT" means its Chinese sentences are acquired by machine translation of the original English sentences. Image captioning models are trained individually on the four training sets and tested exclusively on COCO-CN-test.

Dataset	Usage	Visual instances	Sentences
COCO-CN-dev	training	18,342	20,065
COCO-MT	training	121,286	606,771
VATEX-CN	training	23,896	238,960
VATEX-MT	training	23,896	238,960
COCO-CN-test	test	1,000	6,033

Choice of Network Architecture. We investigate three representative architectures, namely Show and Tell (Show-Tell) [24], Bottom-up and Top-Down (Up-Down) [2] and Attention on Attention Network (AoANet) [9]:

- **Show-Tell**: Proposed by Vinyals *et al.* [24], this model generates a caption for a given image in an encoding-decoding manner. The given image is encoded as a feature vector by a pre-trained image CNN model. The feature vector is then used as an input of an LSTM network which iteratively generates a sequence of words as the generated caption.
- **Up-Down**: Proposed by Anderson *et al.* [2], this model improves Show-Tell by introducing a combined bottom-up and top-down visual attention mechanism. In contrast to the global feature used in Show-Tell, Up-Down encodes the given image by a varied number of feature vectors, extracted from objects detected by Faster R-CNN. Such a design not only describes dominant patterns in the image but also captures small-sized objects. In the decoding stage, a weighted average of these features is fed into an LSTM network, with the weights calculated by a self-attention module to adaptively reflect the importance of the individual features for caption generation. In this work, we use visual features from [19].
- **AoANet**: Proposed by Huang *et al.* [9], this model improves the previous Up-Down model by introducing an Attention on Attention (AoA) module. AoA extends the conventional attention mechanism by adding a second attention layer, allowing the module to take into account the relevance between the query vector (which is the input of the attention module) and the attention result. AoANet is built by applying AoA to Up-Down's encoder and the decoder.

Given the four datasets and the three networks, we shall have 12 models in total. However, as classes and positions of the detected objects vary over frames, Up-Down and AoANet are not directly applicable to video data. Hence, only Show-Tell is trained on all the four datasets. This results in 8 distinct models, see Table 2. Each model is named after the underlying network and training data. E.g., AoANet (COCO-MT) means training AoANet on COCO-MT.

4.1.2 Details of Implementation. All the image captioning models are trained in a standard supervised manner, with the cross-entropy loss minimized by the Adam optimizer. The initial learning rate of Show-Tell and Up-Down is set to be 0.0005. All hyper-parameters of AoANet follow the original paper [9]. The maximum number of training epochs is 80. Best models are selected based on their CIDEr scores on the validation set of the corresponding dataset.

All models are exclusively tested on the test set of COCO-CN, which has 1,000 images. Each test image is associated with five English sentences originally provided by MS-COCO and on average six Chinese sentences. We use the first English sentence as y_s.

The English version of W2VV++ is trained on paired image and English captions from MS-COCO, with 121k images and 606k captions in total. Note that the images have no overlap with the test set. As for the Chinese version of W2VV++, we pretrain the model using COCO-MT and fine-tune it on COCO-CN-dev. Given the relatively limited availability of bilingually annotated image data, how to train the model in a semi-supervised manner [27] deserves further investigation.

4.2 Experiment 1. Evaluation of the Proposed Metrics in Scenario-I

We summarize the performance of the eight models measured by the varied metrics in Table 2, where BMRC is the sum of BLEU-4, METEOR, ROUGE-L and CIDEr, while WCC is the sum of WMDRel, CLinRel and CMedRel. According to both CIDEr and BMRC, AoANet (COCO-MT) has the top performance, while models using the bottom-up and top-down visual features outperform their Up-Down counterparts. This result is reasonable, in line with the literature that attention mechanisms are helpful. We observe Table 2 that such a model preference is also identified by WCC.

Table 2: Performance of distinct models for generating Chinese captions, measured by standard and proposed metrics. BMRC is the sum of BLEU-4, METEOR, ROUGE-L and CIDEr, while WCC is the sum of WMDRel, CLinRel and CMedRel. Models sorted in descending order by BMRC. Both BMRC and WCC rank AoANet (COCO-MT) as the top-performing model.

	Standard Metrics					Proposed Metrics			
Model	*BLEU-4*	*METEOR*	*ROUGE-L*	*CIDEr*	*BMRC*	*WMDRel*	*CLinRel*	*CMedRel*	*WCC*
AoANet (COCO-MT)	33.5	**29.4**	52.7	**97.5**	**213.1**	51.1	**42.7**	**33.5**	**127.3**
Up-Down (COCO-CN)	**36.1**	28.7	**54.3**	92.2	211.3	53.3	37.8	32.2	123.3
AoANet (COCO-CN)	34.4	29.2	53.8	92.3	209.7	**53.6**	39.4	33.4	126.4
Up-Down (COCO-MT)	31.8	27.9	51.0	91.0	201.7	49.8	39.8	31.5	121.1
Show-Tell (COCO-CN)	32.3	27.2	51.8	85.1	196.4	52.1	34.7	32.1	118.9
Show-Tell (COCO-MT)	30.6	27.2	50.3	87.0	195.1	49.4	39.0	32.6	121.0
Show-Tell (VATEX-MT)	12.0	20.0	35.5	34.3	101.8	40.4	1.0	23.0	64.3
Show-Tell (VATEX-CN)	9.9	20.9	35.1	29.1	95.0	40.6	1.9	20.6	63.1

Comparing the individual models, Up-Down (COCO-CN) obtains a higher BMRC than AoANet (COCO-CN), although [9] reports that AoANet is better than Up-Down for English image captioning on MS-COCO. Meanwhile, we notice that AoANet (COCO-MT) has a higher BMRC than Up-Down (COCO-MT). Recalling that the amount of training sentences in COCO-MT is around 30 times as large as that of COCO-CN. Hence, the advantage of AoANet is subject to the amount of training data.

Also noticing that models trained on COCO-CN obtain higher BLEU-4 than their counterparts trained on COCO-MT. We attribute this result to the reason that the COCO-CN models generate longer sentences, while BLEU-4 adds a brevity-penalty to discourage short sentences. As CIDEr does not take the length of a sentence into account, this explains why some image captioning models have higher CIDEr yet lower BLEU-4.

The effectiveness of the proposed metrics is justified by the Spearman correlation reported in Table 3. Among them, WMDRel is most correlated with BLEU-4, CLinRel with CIDEr, and CMedRel with CIDEr. We also evaluate varied combinations of the proposed metrics. Among them, WCC has the largest Spearman correlation of 1.0 with CIDEr and 0.952 with BMRC. Thus, WMDRel, CLinRel and CMedRel shall be used together for Scenario-I.

4.3 Experiment 2. Evaluation of the Proposed Metrics in Scenario-II

As aforementioned, only CMedRel is applicable in Scenario-II, which is much more difficult by definition. As shown in Table 3, the Spearman correlation coefficients of CMedRel with BLEU-4, METEOR, ROUGH-L, CIDEr and BMRC are 0.714, 0.838, 0.714, 0.881, and 0.786, respectively. All the coefficients are greater than 0.7. This result indicates that CMedRel has good correlations with the standard metrics. Hence, the metric can be used with caution when no reference sentence is available.

For a more intuitive understanding of the results, some generated captions and the corresponding metrics computed upon these captions are presented in Table 4.

Table 3: Spearman's rank correlation coefficient between model ranks are produced by the proposed metrics and by the standard metrics separately. The bold number in each column highlights one of the proposed metrics that is most correlated to a standard metric. A coefficient of 1 means identical model ranks.

Proposed Metric	**BLEU-4**	**METEOR**	**ROUGE-L**	**CIDEr**	**BMRC**
WMDRel	0.929	0.778	0.929	0.714	0.762
CLinRel	0.524	0.862	0.524	0.857	0.762
CMedRel	0.714	0.838	0.714	0.881	0.786
WMD + CLin	0.810	**0.994**	0.810	0.976	0.929
WMD + CMed	**0.952**	0.826	**0.952**	0.833	0.833
CLin + CMed	0.595	0.850	0.595	0.905	0.762
WCC	0.833	0.970	0.833	**1.000**	**0.952**

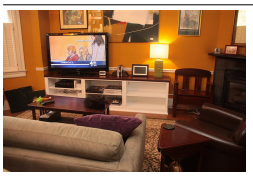
5 CONCLUSIONS AND REMARKS

This paper presents our effort towards annotation-free evaluation of cross-lingual image captioning. Experiments on two cross-lingual datasets (COCO-CN and VATEX) and three representative image captioning networks (Show-Tell, Up-Down and AoANet) allow us to draw conclusions as follows. When each test image is associated with one reference sentence in the source language, the combination of the three proposed metrics (WMDRel, CLinRel and CMedRel) has perfect Spearman correlation of 1 with CIDEr and 0.952 with BMRC. When such cross-lingual references are unavailable, CMedRel still has Spearman correlation of 0.881 with CIDEr and 0.786 with BMRC. These results suggest that the current need of references in the target language can be largely reduced. This will enable a more effective utlization of expensive and thus limited human resources on assessing subjective properties, *e.g.,* readability and fluency, of the auto-generated captions.

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. *Adaptive Behavior* 11, 4 (2016), 382–398.

Table 4: Examples of automatically generated Chinese captions and their quality measured by distinct metrics. For each test image shown in this table, the generated captions are sorted in descending order in terms of BMRC. Texts in parentheses are English translations provided for non-Chinese readers. Due to the domain gap, models trained on VATEX-CN / VATEX-MT are less effective than their counterparts trained on COCO-CN / COCO-MT. This is confirmed by the relatively lower scores of the proposed metrics.

Test image	Generated caption $\hat{y}_t$	CIDEr	BMRC	WMDRel	CLinRel	CMedRel	WCC
 y_s: of a clock on the top of a building $MT(y_s)$: 建筑物顶部的时钟	**Up-Down(COCO-MT)**: 上面有一个钟的大建筑物 (A large building with a clock on it)	100.8	195.0	51.7	54.5	32.7	138.9
	Show-Tell(COCO-MT): 上面有一个钟的大建筑物 (A large building with a clock on it)	100.8	195.0	51.7	54.5	32.7	138.9
	AoANet(COCO-MT): 有一个钟的大建筑物 (A large building with a clock)	99.7	184.1	47.9	55.1	33.2	136.2
	Up-Down(COCO-CN): 一个古老的建筑物上有一个钟 (There is a clock on an old building)	71.8	157.4	49.7	49.3	29.6	128.6
	AoANet(COCO-CN): 一座古老的建筑物上有一个钟 (There is a clock on an old building)	73.2	153.4	61.9	49.1	34.3	145.3
	Show-Tell(COCO-CN): 一座古老的教堂 (An old church)	6.2	58.1	37.9	15.0	20.8	73.7
	Show-Tell(VATEX-CN): 一个穿着黑色衣服的人正在房间里玩 (A man in black is playing in the room)	0.3	39.8	38.2	-2.2	0.2	36.2
	Show-Tell(VATEX-MT): 一个人正在用一种特殊的工具在墙上画 (A man is painting on the wall with a special tool)	0.2	39.5	41.8	-6.5	7.0	42.3
	Spearman's rank correlation with CIDEr	-	-	0.744	0.915	0.783	0.851
	Spearman's rank correlation with BMRC	-	-	0.680	0.936	0.695	0.979
 y_s: couple of women standing next to a red stop sign $MT(y_s)$: 几个女人站在一个红色的停车标志旁边	**AoANet(COCO-CN)**: 两个女人站在停车标志旁 (Two women standing by the stop sign)	216.1	442.4	89.0	75.0	47.7	211.7
	Up-Down(COCO-MT): 一个男人和一个女人站在停车标志旁边 (A man and a woman standing next to the stop sign)	156.5	327.5	70.7	65.1	50.9	186.7
	AoANet(COCO-MT): 两个人站在停车标志旁边 (Two people standing next to the stop sign)	121.8	291.5	78.0	64.9	50.9	193.8
	Show-Tell(COCO-MT): 三个人站在停车标志旁边 (Three people standing next to the stop sign)	105.0	253.5	73.9	64.0	53.3	191.2
	Up-Down(COCO-CN): 一个男人和一个女人站在街道旁 (A man and a woman standing by the street)	60.0	139.2	67.7	43.5	25.9	137.1
	Show-Tell(VATEX-MT): 两个人坐在一张桌子旁, 其中一个人在着一只鞋 (Two people sitting at a table, one of them is wearing a shoe)	13.2	76.8	48.4	2.2	20.6	71.2
	Show-Tell(VATEX-CN): 两个穿着红色衣服的男人正坐在一起 (Two men in red are sitting together)	19.9	74.0	45.8	-2.4	14.9	58.3
	Show-Tell(COCO-CN): 一个穿着红色衣服的女人在街道上打电话 (A woman in red is calling on the street)	22.5	68.7	38.6	39.5	16.4	94.5
	Spearman's rank correlation with CIDEr	-	-	0.833	0.976	0.718	0.905
	Spearman's rank correlation with BMRC	-	-	0.929	0.929	0.763	0.857
 y_s: is a living room with a couch and television $MT(y_s)$: 是一个有沙发和电视的客厅	**Show-Tell(COCO-MT)**: 有沙发和电视的客厅 (A living room with sofa and TV	155.4	317.8	50.3	79.0	52.5	181.8
	AoANet(COCO-MT): 有沙发椅和电视的客厅 (A living room with sofa chairs and TV)	113.2	251.5	75.7	79.0	52.5	207.2
	Up-Down(COCO-MT): 有沙发椅和电视的客厅 (A living room with sofa chairs and TV)	113.2	251.5	75.7	79.0	52.5	207.2
	Show-Tell(COCO-CN): 客厅里有沙发茶几和电视机 (There is a sofa, tea table and TV in the living room)	120.5	220.3	60.2	70.5	45.6	176.3
	Up-Down(COCO-CN): 客厅里有一台电视和一台电视 (There is a TV and a TV in the living room)	50.5	132.3	43.3	57.5	47.8	148.6
	AoANet(COCO-CN): 客厅里有一台电视 (There is a TV in the living room)	59.9	122.8	37.5	57.7	47.2	142.4
	Show-Tell(VATEX-CN): 一个人坐在沙发上看电视 (A man sitting on the sofa watching TV)	67.2	108.8	28.3	-6.1	25.7	47.9
	Show-Tell(VATEX-MT): 一个穿着黑色衣服的人正在看电视 (A man in black is watching TV)	19.9	76.0	23.1	-2.8	43.3	63.6
	Spearman's rank correlation with CIDEr	-	-	0.717	0.773	0.573	0.674
	Spearman's rank correlation with BMRC	-	-	0.872	0.952	0.913	0.915

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*.

[3] Xinlei Chen, Fang Hao, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2014. Microsoft COCO captions: Data collection and evaluation server. *CoRR* (2014).

[4] Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *SMT*.

[5] Jianfeng Dong, Xirong Li, and Cees G. M. Snoek. 2018. Predicting Visual Features from Text for Image and Video Caption Retrieval. *T-MM* 20, 12 (2018), 3377–3388.

[6] Jiahui Gao, Yi Zhou, Philip L. H. Yu, and Jiuxiang Gu. 2020. Unsupervised Cross-lingual Image Captioning. (2020). arXiv:cs.CL/2010.01288

[7] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. 2018. Unpaired Image Captioning by Language Pivoting. In *ECCV*.

[8] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2015. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research* 47, 1 (2015), 853–899.

[9] Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. 2019. Attention on Attention for Image Captioning. In *CVPR*.

[10] Zhengxiong Jia and Xirong Li. 2020. iCap: Interactive Image Captioning with Predictive Text. In *ICMR*.

[11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. In *CVPR*.

[12] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating Automatic Metrics for Image Captioning. In *EACL*.

[13] Todd Ward Kishore Papineni, Salim Roukos and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

[14] Weiyu Lan, Xirong Li, and Jianfeng Dong. 2017. Fluency-Guided Cross-Lingual ImageCaptioning. In *ACMMM*.

[15] Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 2016. Adding Chinese Captions to Image. In *ICMR*.

[16] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. COCO-CN for Cross-Lingual Image Tagging, Captioning and Retrieval. *T-MM* 21, 9 (2019), 2347–2360.

[17] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. 2019. W2VV++: Fully Deep Learning for Ad-hoc Video Search. In *ACMMM*.

[18] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1013

[19] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *CVPR*.

[20] Pascal Mettes, Dennis C Koelma, and Cees G M Snoek. 2020. Shuffled ImageNet Banks for Video Event Detection and Search. *TOMM* 16, 2 (2020), 1–21.

[21] Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-Lingual Image Caption Generation. In *ACL*.

[22] Yuqing Song, Shizhe Chen, Yida Zhao, and Qin Jin. Unpaired Cross-lingual Image Caption Generation with Self-Supervised Rewards. In *ACMMM*.

[23] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*.

[24] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *CVPR*.

[25] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan Fang Wang, and William Yang Wang. 2019. VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *ICCV*.

[26] Yike Wu, Shiwan Zhao, Jia Chen, Ying Zhang, and Zhong Su. 2019. Improving Captioning for Low-Resource Languages by Cycle Consistency. In *ICME*.

[27] Xun Yang, Meng Wang, Richang Hong, Qi Tian, and Yong Rui. 2017. Enhancing Person Re-Identification in a Self-Trained Subspace. *ACM TOMM* 13, 3, Article 27 (2017), 23 pages.

[28] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* (2014).