Adding Chinese Captions to Images

Xirong Li[†] Weiyu Lan[†] Jianfeng Dong[‡] Hailong Liu[§]
[†]Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China
[‡]College of Computer Science and Technology, Zhejiang University
[§]Pattern Recognition Center, WeChat Department, Tencent

ABSTRACT

This paper extends research on automated image captioning in the dimension of language, studying how to generate Chinese sentence descriptions for unlabeled images. To evaluate image captioning in this novel context, we present Flickr8k-CN, a bilingual extension of the popular Flickr8k set. The new multimedia dataset can be used to quantitatively assess the performance of Chinese captioning and English-Chinese machine translation. The possibility of re-using existing English data and models via machine translation is investigated. Our study reveals to some extent that a computer can master two distinct languages, English and Chinese, at a similar level for describing the visual world. Data is publicly available at http://tinyurl.com/flickr8kcn.

Keywords

Image captioning, Bilingual dataset, Chinese language

1. INTRODUCTION

This paper studies Image Captioning – automatically generating a natural language description for a given image. Different from the ongoing research that focuses on improving computational models for image captioning [2, 6, 12], which is undoubtedly important, we consider a new dimension of the problem: *Language*. While current works are on English, we study the possibility of captioning images in another language. In particular, we consider *Chinese*, the most spoken language in the world.

Extending along the language dimension is not just engineering efforts such as constructing a dataset in another language and applying existing models. We argue that this line of research is scientifically interesting in multiple aspects. An image caption is intended to provide clear descriptions of salient entities and events present in the image. The perception of saliency, however, could be culture dependent. Consider for instance the image shown in the first row of Table 1. While both English and Chinese annotators capture the event of taking a photo, they use distinct qualifiers for the word 'woman': 'Asian' in the English sentence and 中年 (middle-aged) in the Chinese sentence. Asian faces are probably too common to be visually salient from a Chinese point of view. By contrast, in the Chinese caption of the second image, the little girl is qualified with 金色头发 (blonde-haired). Research on multilingual captioning helps reveal how people from culturally and linguistically diverse backgrounds describe the visual world. Further, the divergence may provide complementary information to each of the languages and thus mutually boost the performance of individual language models. More fundamentally, while it remains controversial whether English or Chinese is the harder language to learn for human, it will be interesting to investigate if a computer can master the two languages at the same level for describing images. Towards answering the above questions, image captioning in a bilingual setting is a good starting point.

To build a Chinese-captioning model, an intuitive and inexpensive solution is to employ machine translation. Either in an early stage to automatically translate all the training text from English to Chinese, or in a late stage to translate the output of a pretrained English model. While the use of web-scale data has substantially improved translation quality, machine translation remains unreliable. A relatively simple sentence as 'A young girl in pigtails plays in the water' is translated to 一位年轻的女孩辫子起着水 by Google and 在发挥水的辫子姑娘 by Baidu. Neither of the translation makes sense. Given translation unreliability, questions arise as is machine translation still usable, which translation to use, and in what stage?

In this paper we present a pilot study for Chinese captioning of images. We extending the widely used Flickr8k dataset [5] to a bilingual version, with both machine translated and human written Chinese sentences. The new dataset can be used to assess the accuracy of Chinese captioning and the quality of machine translation. To the best of our knowledge, datasets of this kind do not exist in the public literature. We investigate the possibility of re-using existing English captions and models via machine translation. Three solutions are proposed and evaluated on the new bilingual dataset. Consequently, we identify good strategies for building a Chinese-captioning model.

2. PROGRESS ON IMAGE CAPTIONING

The state-of-the-art follows an image-encoding and sentence-decoding pipeline. A popular implementation, as developed in [3, 6, 12], is to encode an image with a pretrained deep

ICMR'16, June 06 - 09, 2016, New York, NY, USA

ISBN 978-1-4503-4359-6/16/06…$15.00
DOI: http://dx.doi.org/10.1145/2911996.2912049

Table 1: Some examples of Flickr8K-CN, a bilingual multimedia dataset for image captioning.

Image	English caption	Google translation	Baidu translation	Human translation	Chinese caption
	An Asian woman is taking a photograph outside a white columned building	一个亚洲女子走的是一条白色的圆柱状建筑外的照片	一个亚洲女人是白色圆柱的大楼外拍照	一个亚洲女人在一个白色圆柱大楼外拍照	一个中年女人正在拿着照相机准备拍照
	The little girl is running and laughing	这个小女孩奔跑着，欢笑着	小女孩在奔跑和欢笑	小女孩奔跑着，笑着	金色头发的小女孩
	A furry black and white dog jumps over a bar during an agility test	一个毛茸茸的黑色和白色的狗跳在一个酒吧的敏捷性测试中	一个毛茸茸的黑色和白色的敏捷测试过程中在一个酒吧的狗跳	一只毛茸茸的黑白相间的狗在敏捷性测试中跨栏	一只狗在跨栏
	A group of basketball players wearing yellow and green reach for a ball	A组篮球运动员身穿黄色和绿色端起一球	一群穿着黄色和绿色的篮球运动员	一群穿着黄色和绿色球服的篮球运动员在抢球	篮球比赛的现场

Convolutional Neural Networks (CNN), and then feed the image embedding to a Recurrent Neural Network (RNN), which eventually outputs a sequence of words as the caption. As for the RNN architecture, a bidirectional RNN is adopted in [6], while Long Short-Term Memory (LSTM) is utilized in [12]. Different from a one-layer word embedding [6, 12], Mao *et al.* incorporate a two-layer word embedding in their network [3]. LSTM based solutions have been ranked top in recent benchmark evaluations [1, 4, 7].

Besides the good progress on models, novel datasets are continuously developed, e.g., Flickr8k [5], Flickr30k [13] and MSCOCO [8], increasing the number of captioned images from a few thousands to hundreds of thousands. However, all the text is English. More recently, a bilingual dataset on the base of MSCOCO is introduced for answering questions about the content of an image [3]. A sample question-answer pair is 'what is there in yellow?' and 'Banana'. Questions are not captions. They are written in a different motivation. Datasets suited for studying image captioning in a bilingual setting remains to be established.

3. A BILINGUAL MULTIMEDIA DATASET

For captioning images in Chinese, we need a set of images accompanied with Chinese descriptions. For cross-language analysis, we choose to depart from an existing dataset wherein each image is already associated with some English captions. In particular, we employ Flickr8k [5], as it is well recognized in the literature [6, 9, 12] and its relatively small size is more suited for a pilot study. Given the data partition from [6], Flickr8k consists of 8,000 images, where 6,000 images are used for training, 1,000 images for validation, and the remaining 1,000 images for test. Each image has five English sentences collected via Amazon's Mechanical Turk. Written by US residents, these sentences were meant for briefly describing main objects and scenes in the image.

To make Flickr8k bilingual, a straightforward solution is to translate each sentence from English to Chinese by machine translation. We employ English-Chinese translation services provided by Google and Baidu, respectively. Some examples are given in Table 1. We observe that machine translation does not perform well as sentences become longer and contain ambiguous words.

Even though the translated sentences would have been perfect, they do not necessarily reflect how a Chinese describes the same image. In order to acquire sentences that better match with Chinese conventions, we recruited a local crowdsourcing service. The annotation was performed by a number of native speakers of Chinese, where they were asked to write sentences describing salient objects and scenes in every image, from their own point of views. Apart from this, we do not specify any rules on wording, intending to gather sentences written in a free style. Each image in Flickr8k receives five manually written Chinese sentences.

We compare sentences written in the two languages in terms of the most common words, see Table 2. For the ease of analysis, each word is manually assigned to one of four visual classes, i.e., objects, scenes, actions and colors. A noticeable difference is that the Chinese sentences contain much less words related to colors. For instance, the word 'black' is used 3,832 times in the English sentences, while the corresponding Chinese word 黑色 appears 116 times only. One reason is that color–object combinations such as 黑狗 (*black dog*) and 黑衣服 (*black clothes*) are valid and common words in Chinese. Such a convention makes the color-related words underestimated. Unsurprisingly, as the sentences were contributed by multiple persons, we observe that distinct words are used to express the same meaning, e.g., 水里 and 水中 to describe something in water, 跑 and 奔跑 for verb 'run' and 玩 and 玩耍 for verb 'play'.

Table 2: The most common words in the bilingual dataset, shown in four classes, i.e., objects, scenes, actions, and colors. Next to each Chinese word is its English translation, provided for non-Chinese readers.

Objects				**Scenes**				**Actions**				**Colors**			
English		Chinese		English		Chinese		English		Chinese		English		Chinese	
word	count	word	count	word	count	word	count	word	count	word	count	word	count	word	count
dog	8,136	人 *person*	7,841	grass	1,622	草地 *grass*	1,532	wearing	3,062	玩 *play*	2,363	white	3,940	黑狗 *black dog*	424
man	7,266	狗 *dog*	6,541	snow	1,492	雪地 *snow*	942	running	2,073	奔跑 *run*	1,739	black	3,832	白狗 *white dog*	126
boy	3,581	男人 *man*	4,564	field	1,280	水中 *in water*	901	playing	2,008	玩耍 *play*	1,726	red	2,672	红色 *red*	121
woman	3,403	小孩 *kid*	3,195	air	1,058	海边 *seaside*	570	standing	1,789	坐在 *sit on*	1,324	brown	2,563	黑色 *black*	116
girl	3,328	女人 *woman*	2,583	beach	1,046	草坪 *lawn*	526	jumping	1,472	骑 *ride*	1,046	blue	2,268	黄狗 *yellow dog*	115
people	2,887	男子 *man*	1,917	street	943	水里 *in water*	512	sitting	1368	站 *stand*	989	green	1,225	白色 *white*	111
dogs	2,125	孩子 *kid*	1,505	outside	791	沙滩 *beach*	425	holding	1324	拿 *hold*	896	yellow	1,213	黄色 *yellow*	65
ball	1,779	自行车 *bicycle*	1,126	pool	691	街道 *street*	412	walking	1,165	穿着 *wear*	742	orange	741	黑 *black*	51
child	1,545	小狗 *puppy*	1,117	wall	556	地上 *on ground*	361	jumps	979	跑 *run*	681	pink	735	粉色 *pink*	43
person	1,542	小女孩 *little girl*	1,079	mountain	554	路上 *on road*	276	runs	925	滑雪 *ski*	655	colorful	217	红 *red*	31

We may learn Chinese models either from the machine translated text or from the human written text. Due to the divergence between speakers of English and speakers of Chinese in describing images, it is unfair to evaluate models learned from machine translation using the crowdsourced ground truth. To faithfully describe English sentences in Chinese, we generate human translation for the test set.

As aforementioned, the test set is comprised of 1,000 images, each associated with five English sentences. Although these sentences are relatively simple in terms of words and structures, properly translating them into Chinese is nontrivial. We hired seven students in our university. Being native speakers of Chinese, they are fluent in English. Examples of human translation are shown in the fifth column of Table 1. On average a human translated sentence contains 15.9 words, almost double the number of a crowdsourced sentence (which is 7.5 words).

Putting all the above together, we have augmented Flickr8k with Chinese sentences generated by machine translation and crowdsourcing, respectively. Moreover, we provide human translation of the 5,000 English sentences in the test set. All this allows us to assess varied approaches to learning Chinese-captioning models.

4. CHINESE-CAPTIONING MODEL

For a novel image I, we aim to automatically predict a Chinese sentence S that describes in brief the visual content of the image. Naturally, the sentence is a sequence of n Chinese words, $S = \{w_1, \ldots, w_n\}$. One way to acquire S is to employ a pre-trained English-captioning model and translate its output from English to Chinese by machine translation. In contrast to such a Late Translation strategy, we discuss in this section how to build a Chinese model directly from a Chinese multimedia corpus, denoted as $\mathcal{C} = \{(I_i, S_{i,1}, \ldots, S_{i,m_i})\}$, where the i-th training image is accompanied with m_i sentences.

In this work we investigate the Neural Image Captioning (NIC) model for generating Chinese sentences, as shown in Fig. 1. NIC is a probabilistic model that uses an LSTM neural network to compute the posterior probability of a sentence given an input image. Consequently, the image will be annotated with the sentence that yields the maximal probability. Given θ as the model parameters, the probability is expressed as $p(S|I;\theta)$. Applying the chain rule together with log probability for the ease of computation, we have

$$\log p(S|I;\theta) = \sum_{t=0}^{n+1} \log p(w_t|I, w_0, \ldots, w_{t-1};\theta), \qquad (1)$$

where $w_0 = \text{START}$ and $w_{n+1} = \text{END}$ are two special tokens indicating the beginning and the end of the sentence.

Conditional probabilities in Eq (1) are estimated by the LSTM network in an iterative manner. The network maintains a cell vector c and a hidden state vector h to adaptively memorize the information fed to it. The embedding vector of an image, obtained by applying an affine transformation on its visual feature vector, is fed to the network to initialize the two memory vectors. In the t-th iteration, new probabilities p_t over each candidate word are re-estimated given the current chosen words. The word with the maximum probability is picked up, and fed to LSTM in the next iteration. The recurrent connections of LSTM carry on previous context. Following [6,12], per iteration we apply beam search to maintain k best candidate sentences. The iteration stops once the END token is selected. To express the above process in a more formal way, we write

$$\begin{aligned} x_{-1} &:= W_e \cdot CNN(I), & (2)\\ x_t &:= W_s \cdot \mathbf{w}_t, \quad t = 0, 1, \ldots, & (3)\\ p_0, c_0, h_0 &\leftarrow \text{LSTM}(x_{-1}, \mathbf{0}, \mathbf{0}), & (4)\\ p_{t+1}, c_{t+1}, h_{t+1} &\leftarrow \text{LSTM}(x_t, c_t, h_t). & (5) \end{aligned}$$

The parameter set θ consists of W_e, W_s, and affine transformations inside LSTM. While DeViSE directly takes a pre-trained word2vec model to construct W_s, NIC optimizes W_e and W_s simultaneously via maximum-likelihood estimation:

$$\max_{\theta} \sum_{(I,S)\in\mathcal{C}} \log p(S|I;\theta). \qquad (6)$$

We use ET-NIC and CS-NIC to indicate models respectively learned from machine translated and crowdsourced Chinese sentences. LT-NIC refers to the use of machine translation in a late stage.

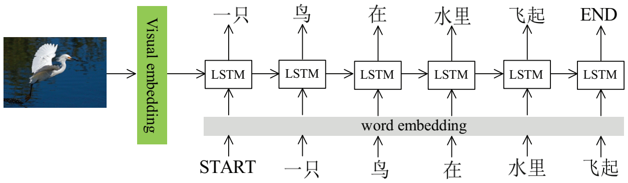

Figure 1: The Chinese-captioning model.

5. EVALUATION

5.1 Setup

Preprocessing. Unlike an English sentence that has spaces as explicit word boundary markers, a Chinese sentence lacks such markers. In order to tokenize a Chinese sentence into a list of meaningful words, we employ Jieba[1], an open-source software for Chinese text segmentation.

For the CNN feature, we employ the pool5 layer of a pre-trained GoogLeNet [11]. For all models the size of the visual and word embeddings is set to be 512.

Concerning the choice of machine translation services, we find that Baidu yields higher BLEU scores than Google for English-to-Chinese and Chinese-to-English translation. So Baidu translation is used in the following experiments.

Evaluation criteria. We adopt BLEU [10], originally developed for automatic evaluation of machine translation and now widely used to evaluate image captioning. To better understand the cross-language influence, we report BLEU with and without the corpus-level brevity penalty

5.2 Experiments

Experiment 1. Which Model for Chinese Captioning? With crowdsourced sentences as references, CS-NIC performs the best, see Table 3. When using human translated sentences as references, ET-NIC is the best, see Table 4. This is largely due to the divergence between crowdsourced sentences and human/machine translated sentences. For instance, as the former is shorter, sentences generated by CS-NIC tend to be shorter than those from ET-NIC and LT-NIC. This also explains why the brevity penalty does not affect BLEU scores of ET-NIC and LT-NIC when evaluating on crowdsourced references. Since machine translated sentences are on average one word shorter than human translation, the penalty causes decreases in BLEU scores when evaluating on human translated references. CS-NIC has a more dramatic decrease. The result also shows the importance of the brevity penalty for cross-corpus evaluation.

As for the two machine translation based models, ET-NIC surpasses LT-NIC. English-to-Chinese translation is error-prone, with BLEU-1 of 65.8 and BLEU-2 of 47.7. The superior performance of ET-NIC implies that the LSTM network has a sort of ability to tolerate the translation error.

Given what we have observed, the answer to the question of which model is more suited for Chinese captioning depends on what ground truth is in use. Nevertheless, we consider ET-NIC more appealing as it performs relatively well and requires no any extra manual annotation.

Experiment 2. Which Language is More Difficult to Learn? For a cross-language comparison, Table 5 shows the performance of the English model. Recall that NIC, ET-NIC and CS-NIC share the same model architecture, except that NIC is trained on English text, while the other two are trained on Chinese text. Moreover, they are tested on the same image set except for the reference sentences. So their comparison gives us a brief idea of how the machine masters each language to describe images.

Table 3: Performance of Chinese-captioning models, using crowdsourced sentences as references.

Model	BLEU-1	BLEU-2	BLEU-3
ET-NIC	45.8 (45.8)	22.8 (22.8)	10.5 (10.5)
LT-NIC	41.1 (41.1)	18.4 (18.4)	7.2 (7.2)
CS-NIC	**63.4** (61.1)	**41.6** (40.1)	**22.1** (21.3)

Table 4: Performance of Chinese-captioning models, using Human translated sentences as references.

Model	BLEU-1	BLEU-2	BLEU-3
ET-NIC	62.9 (53.9)	**37.4** (32.1)	**19.6** (16.8)
LT-NIC	56.8 (46.8)	30.1 (24.8)	13.5 (11.2)
CS-NIC	**63.3** (34.8)	31.3 (17.2)	7.2 (3.9)

Table 5: Performance of English-captioning models.

Model	BLEU-1	BLEU-2	BLEU-3
NIC	**63.4** (57.9)	**40.5** (37.0)	**19.8** (18.1)

For NIC, it obtains BLEU-1 at 57.9 and BLEU-2 at 37.0 (with the brevity penalty). The corresponding numbers for ET-NIC are 53.9 and 32.1, and 61.1 and 40.1 for CS-NIC. The better performance of CS-NIC is presumably because crowdsourced sentences generally contain less words, and shorter text is in principle easier to reconstruct. The relative lower performance of ET-NIC is not only because the length of the reference sentences has doubled, but also because the training text is machine translated (which has BLEU-1 at 62.7 after the brevity penalty). Interestingly, the averaged performance of the two Chinese models, with BLEU-1 at 57.5 and BLUE-2 at 36.1, is on par with the English model. The result suggests that the LSTM network has the capability to learn the two languages at a similar level.

6. CONCLUSIONS

Given image captioning experiments in a bilingual setting, our conclusions are as follows. In spite of its unreliability, machine translation can be used, in particular in an early stage in advance to model learning, for building a Chinese-captioning model. Baidu translation is preferred to Google translation. Which model is more suited for Chinese captioning depends on what ground truth is used. We consider ET-NIC promising as it performs relatively well and requires no extra manual annotation. Our evaluation suggests that the NIC model has the capability to learn both English and Chinese at a similar level for describing images.

Acknowledgements. This research was supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 16XNQ013).

[1] https://github.com/fxsjy/jieba/

7. REFERENCES

[1] MS COCO Captioning Challenge. http://mscoco.org/dataset/#captions-challenge2015.

[2] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt, L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *Proc. of CVPR*, 2015.

[3] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? Dataset and methods for multilingual image question answering. In *Proc. of NIPS*, 2015.

[4] A. Gilbert, L. Piras, J. Wang, F. Yan, E. Dellandrea, R. Gaizauskas, M. Villegas, and K. Mikolajczyk. Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. In *CLEF working notes*, 2015.

[5] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899, 2013.

[6] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. of CVPR*, 2015.

[7] X. Li, Q. Jin, S. Liao, J. Liang, X. He, Y. Huo, W. Lan, B. Xiao, Y. Lu, and J. Xu. RUC-Tencent at ImageCLEF 2015: Concept detection, localization and sentence generation. In *CLEF working notes*, 2015.

[8] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick. Microsoft COCO: Common objects in context. *CoRR*, abs/1405.0312, 2014.

[9] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). In *Proc. of ICLR*, 2015.

[10] K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proc. of ACL*, 2002.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. of CVPR*, 2015.

[12] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proc. of CVPR*, 2015.

[13] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.