The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Generating Image Captions in Arabic Using Root-Word Based Recurrent Neural Networks and Deep Neural Networks

Vasu Jindal
University of Texas at Dallas, Texas, USA
vasu.jindal@utdallas.edu

Abstract

Automatic caption generation of an image requires both computer vision and natural language processing techniques. Despite of advanced research in English caption generation, research on generating Arabic descriptions of an image is extremely limited. Semitic languages like Arabic are heavily influenced by root-words. We leverage this critical dependency of Arabic and in this paper are the first to generate captions of an image directly in Arabic using root-word based Recurrent Neural Networks and Deep Neural Networks. We report the first BLEU score for direct Arabic caption generation. Experimental results confirm that generating image captions using root-words directly in Arabic significantly outperforms the English-Arabic translated captions using state-of-the-art methods.

With the increase in number of devices with cameras, there is a widespread interest in generating automatic captions from images and videos. Automatic generation of image descriptions is a widely researched problem. It has huge impact in the fields of information retrieval, accessibility for the vision impaired, categorization of images etc. Most visual recognition models and approaches in the image caption generation community are focused on Western languages, ignoring Semitic and Middle-Eastern languages like Arabic. This is primarily due to two major reasons: i) Lack of existing image corpora in languages other than English ii) the significant dialects between different forms of Arabic and the challenges in translating images to natural sounding sentences. Given the high influence of Arabic in current political settings, it is necessary for a robust approach for Arabic caption generation.

Semitic languages like Arabic are significantly influenced by their original root-word. We leverage this critical aspect of Arabic to formalize a three stage approach to generate Arabic captions. Our main contribution in this paper is threefold:

- Mapping of image fragments onto root words in Arabic rather than actual sentences or words/fragments of sentences as suggested in previously proposed approaches.
- Finding most appropriate words for an image by using a root-word based Recurrent Neural Networks.

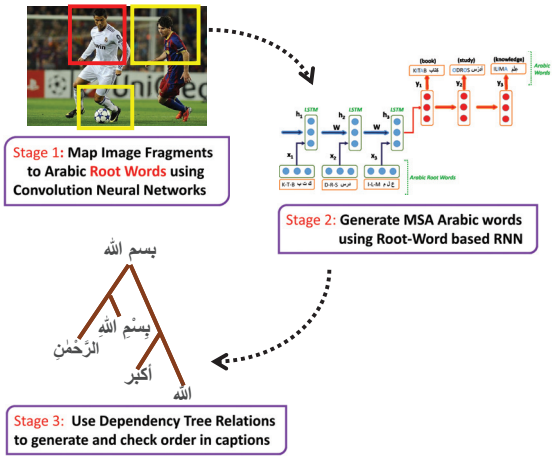

Figure 1: Overview of Our Approach

- Finally, using dependency tree relations of these obtained words to check order in sentences in Arabic.

To the best of our knowledge, this is the first work that leverage root words to generate captions in Arabic. We also report the first BLEU scores for Arabic caption generation. *Additionally, this opens a new field of research to use root-words to generate captions from images in Semitic languages.*

Figure 1 gives a brief overview of our approach. In Stage 1, we map image fragments onto root words in Arabic. We apply the approach (Jia et al. 2014) to detect objects in every image with a Region Convolutional Neural Network (RCNN). It should be noted that the output of the convolutional neural network are Arabic root-words. To achieve this, at any given time when English labels of objects were used in the training of convolution neural networks, Arabic root-words of the object were also provided as input in the training phase. A transducer based algorithm for Arabic root extraction (Yaseen and Hmeidi 2014) is used to extract root-words from an Arabic word in the training stage.

In Stage 2, we used root word based Recurrent Neural Networks with LSTM memory cell to generate the most appropriate words for an image in Modern Standard Arabic

Dataset	Model	BLEU				METEOR
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	
Flickr8k	BRNN (Karpathy and Fei-Fei 2015)	48.2	45.1	29.4	15.5	—
	Google (Vinyals et al. 2015)	52.4	46.1	34.8	18.6	—
	Ours	**65.8**	**55.9**	**40.4**	**22.3**	**20.09**
Middle Eastern News Websites	BRNN (Karpathy and Fei-Fei 2015)	45.4	34.8	27.6	13.9	12.11
	Google (Vinyals et al. 2015)	46.2	36.4	28.5	14.3	15.18
	Ours	**55.6**	**43.3**	**34.5**	**18.9**	**18.01**

Table 1: BLEU-1,2,3,4/METEOR metrics compared to other methods, (—) indicates an unknown metric

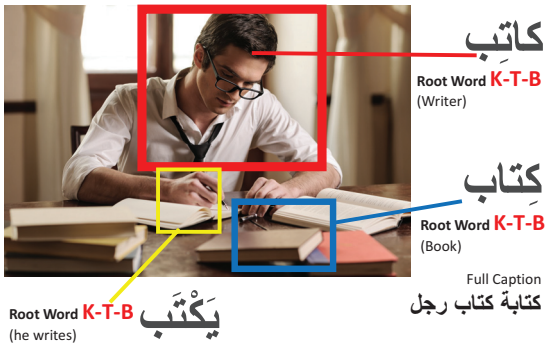

Figure 2: State-of-the-Art: Man studying with books
Ours (translated from Arabic for reader's readability): Writer writing on notebook

(MSA). We propose a root-word based recurrent neural network (rwRNN). The model takes different root-words extracted from text, and predicts the most appropriate words for captions in Arabic. While a standard input vector for RNN derives from either a word or a character, the input vector in rwRNN consists of a root-word specified with 3 letters (r_{1n}, r_{2n}, r_{3n}) that correspond to the characters in root-words' position. Most root-words in Arabic are trilateral and very few are quadilateral or pentalateral. If a particular root-word is quadilateral (pentalateral) then the r_{2n} represents the middle three (four) letters of the root-word. Formally:

$$x_n = \begin{bmatrix} r_{1n} \\ r_{2n} \\ r_{3n} \end{bmatrix} \quad (1)$$

The final output (i.e. predicted actual Arabic word y_n), the hidden state vector (h_n) of the LSTM is taken as input to the softmax function layer with a fixed vocabulary size. Dependency tree constraints checks the caption generated from RNN to be grammatically valid in Modern Standard Arabic and robust to different diacritics in Arabic. The most popular Prague Arabic Dependency Treebank (PADT) consisting of multi-level linguistic annotations over Modern Standard Arabic is used for the dependency tree constraints (Hajic et al. 2004). Figure 2 shows the result of our approach for a sample image.

Experimental Results

We evaluate our technique using two datasets: Flickr8k dataset with manually written captions in Arabic by professional Arabic translators and 405,000 images with captions from various Middle Eastern countries' newspapers. All results shown in Table 1 are captions generated using the corresponding approaches in English and translating them to Arabic using Google Translate. According to Table 1, we can see that our root-word based approach outperforms all current English based approaches and translated to Arabic using Google Translate. *The results also show that generating captions directly in Arabic attains a much better BLEU scores rather than generating captions in English and translating them to Arabic.*

Conclusion and Future Work

This paper presents a novel three-stage technique for automatic image caption generation using a combination of root-word based recurrent neural network and root-word based deep convolution neural network. This is the first reported BLEU score for direct Arabic caption generation and the experimental results show a promising performance.

References

Hajic, J.; Smrz, O.; Zemánek, P.; Šnaidauf, J.; and Beška, E. 2004. Prague arabic dependency treebank: Development in data and tools. In *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, 110–117.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, 675–678. ACM.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.

Yaseen, Q., and Hmeidi, I. 2014. Extracting the roots of arabic words without removing affixes. *Journal of Information Science* 40(3):376–385.