CUTIE: Learning to Understand Documents with Convolutional Universal Text Information Extractor

Xiaohui Zhao, Zhuo Wu, and Xiaoguang Wang
New IT Accenture
xiaohui.zhao@accenture.com zhuo.wu@accenture.com danny.x.wang@accenture.com

Abstract

Extracting key information from documents, such as receipts or invoices, and preserving the interested texts to structured data is crucial in the document-intensive streamline processes of office automation in areas that includes but not limited to accounting, financial, and taxation areas. Large proportion of published works attempt to tackle the problem by exploring the semantic context in text sequences based on the named entity recognition method in NLP field. In this paper, we propose to combine the effective information from both semantic meaning and spatial distribution of texts in documents. Specifically, our proposed model, Convolutional Universal Text Information Extractor (CUTIE), applies convolutional networks on the gridded texts where texts are embedded as features with semantical connotations. We further explore the effect of employing different structures of convolutional network and propose a faster and portable structure. We demonstrate the effectiveness of the proposed method on a dataset with up to 3000 labelled receipts, without any post-processing, achieving state-of-the-art performance that is higher than BERT but with only 1/10 parameters and much higher than the RNN based methods.

1. Introduction

Implementing Scanned receipts OCR and information extraction (SROIE) is of great benefit to services and applications such as efficient archiving, compliance check, and fast indexing in the document-intensive streamline processes of office automation in areas that includes but not limit to accounting, financial, and taxation areas. There are two specific tasks involved in SROIE: receipt OCR and key information extraction. In this work, we focus on the second task that is rare in published research. In fact, key information extraction faces big chanllenges, where different types of document structures and vast number of potential interested key words introduces great difficulties. Although the commonly used rule-based method can be implemented with carefully designed expert rules, it can only work on specific type of documents and takes no lesser effort to adapt to new type of documents. Therefore, it is desirable to have a learning based key information extraction method with limited requirement of human resources and solely employing the deep learning technique without designing expert rules for any specific type of documents.

History about published learning based methods goes here. The majority of the published learning based works are based on the named entity recognition research in the NLP field that is not originally designed to solve the key information extraction problem in SROIE, which align text words in the original document as a long paragraph in line-based rule. However, the real world documents, such as receipts and invoices, present with various styles of layouts that were designed for different scenarios or from different enterprise entities, which introduces great difficulties. The order or distance of the texts in the line-based aligned long paragraph tend to vary greatly due to layout variations, which is difficult to be handled with the natural language paragraph oriented methods, one typical example is illustrated in Fig. 2.

Attempting to involve the spatial information into the key information extraction process, we propose to tackle this problem by using the CNN based network structure and involve the semantic features in a properly designed fashion. In particular, our proposed model, called Convolutional Universal Text Information Extractor (CUTIE), tackles the key information extraction problem by applying convolutional deep learning model on the gridded texts, as illustrated in Fig. 1. The gridded texts is the formatted input of the entire network, which is arranged in close correspondance with their real spatial relationship of the original scanned document image. Furthermore, the rich semantic information is encoded from the gridded texts in the very beginning of the convolutional network.

2. Related Work

3. Method

3.1. Grid Positional Mapping

To generate input grid data for the convolutional network, the scanned document image are processed by an OCR engine to acquire the texts and their absolute / relative positions. Let the minimum bounding box around an interested text s be b_s that is restricted by four corner coordinates, where the upper-left corner coordinate in the scanned document be (x_{left}, y_{top}) and the bottom right of the bounding box be (x_{right}, y_{bottom}). To avoid the affects from overlapped bounding boxes and reveal the actual relative position among texts, we calculate the center point (c_x, c_y) of the bounding boxes as the reference position. It is not hard to find that involving pre-processes that combine texts into meaningful entities here will benefit the following grid positional mapping process. However, this is not the major purpose of this paper and we leave it to future researches. In this paper, we tokenize the text words with a greedy longest-match-first algorithm using a pre-defined dictionary [1].

Let the grid positional mapping process be G and the target grid size be (c_{g_0}, r_{g_0}). To generate the grid data, the goal of G is to map the texts from the original scanned document image to the target grid, such that the mapped grid preserves the original spatial relationship among texts yet more suitable to be used as the input for the convolutional network.

3.2. Model

Through matching the output of CUTIE with the labelled grid data, the model learns to generate the label for each text in the grid input via exploring both the spatial and semantic features. For that reason, the task of CUTIE holds resemblance to the semantic segmentation task in the computer vision field. Several methods have been proposed to capture multi-scale context in the input data. The encoder-decoder structure. We propose a convolutional network

$$c_h = x_l + \frac{(x_r - x_l)}{2} \tag{1}$$

$$c_v = y_t + \frac{(y_b - y_t)}{2} \tag{2}$$

For tokenized texts, the bounding box is horizontally divided into multiple boxes and their reference positions are calculated using the same criteria as Equ. 1 and Equ. 2, separately.

Figure 1. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

3.3. Grid Augmentation

4. Experiments

5. Discussion

Automatically extracting interested words / information from the scanned document images is of great interest to various services and applications. The performance gain is mainly achieved by exploring three key factors: the spatial relationship among texts, the semantic information of texts, and the text gridding mechanism.

5.1. Miscellaneous

Compare the following:

`$conf_a$`	$conf_a$
`$\mathit{conf}_a$`	$\mathit{conf}_a$

See The TeXbook, p165.

The space after *e.g.*, meaning “for example”, should not be a sentence-ending space. So *e.g.* is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

When citing a multi-author paper, you may save space by using “et alia”, shortened to “*et al.*” (not “*et. al.*” as “*et*” is a complete word.) However, use it only when there are three or more authors. Thus, the following is correct: “ Frobnication has been trendy lately. It was introduced by Alpher [**?**], and subsequently developed by Alpher and Fotheringham-Smythe [**?**], and Alpher *et al.* [**?**].”

This is incorrect: “... subsequently developed by Alpher *et al.* [**?**] ...” because reference [**?**] has just two authors. If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al.*

References

[1] Google. pre-trained models for bert.

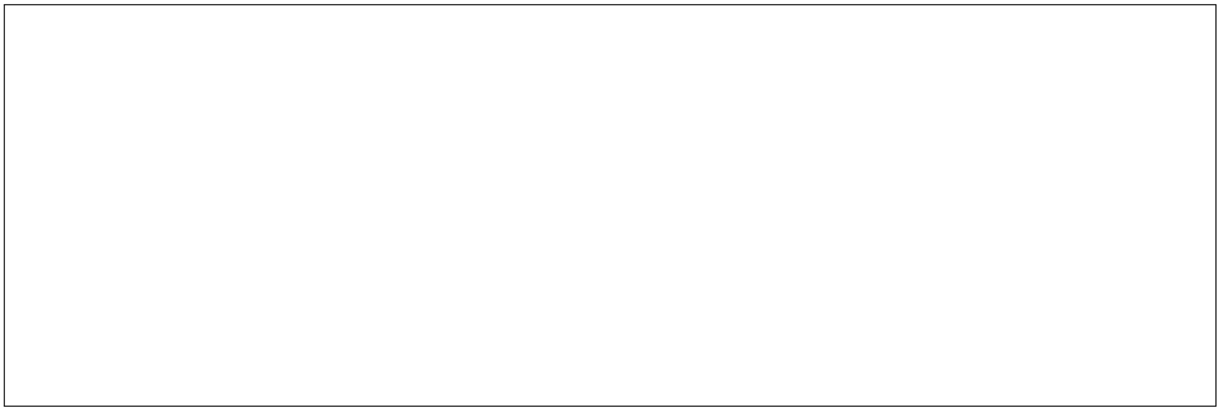

Figure 2. Example of a short caption, which should be centered.