Recognition of Invoices from Scanned Documents

Hien Thi Ha

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanicka 68a, 602 00, Brno, Czech Republic
xha1@fi.muni.cz

Abstract. In this paper, we describe the work of recognition the first page of an invoice from a set of scanned business documents. This can be applied to document management systems, document analysis systems, pre-processing of information extraction systems. We also present our experiments on Czech and English invoice data set.

Key words: classification, recognition, invoice, OCR, Czech

1 Introduction

The processing of business documents, particularly invoices is playing a vital role in companies, especially in large ones. Usually, invoices are classified and relevant details are extracted manually by staff members and input into database systems for further processing. This manual process is time-consuming, expensive and at risk of errors because of large volumes, various layouts and different delivery formats. For those reasons, automatic analysis systems become essential.

An overview of document analysis and recognition systems can be found in [4]. State-of-the-art of text classification algorithms is in [5], [6], [7]. In these papers, authors list different approaches for text classification tasks. Some argue that Support Vector Machine is more suitable for text data while others state that Naive Bayes is more effective ([6]). In general, bag of word is the most popular method to represent document but features' dimension is huge. Therefore, they pay more attention on feature selection techniques. Fortunately, invoices do not have so many words in common (items are not taken account into these shared keywords). Moreover, scanned invoices have typical layout structures such as blocks and tables. In [2] and [8], they presented rule based approach and case-bases reasoning method for document structure recognition. Furthermore, information extraction from invoices are proposed in [3], [9]. There is a note that, in these systems, they used commercial OCR systems to process invoice images.

In this paper, we focus on classifying invoices of Czech companies which are mainly in Czech and English. The paper is constructed as follows: in section 2, we describe the classification task. Then, we will explain our experiments and discuss the results gained from our data set in section 3 and finally are conclusion and future work in section 4.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2017, pp. 71–78, 2017.

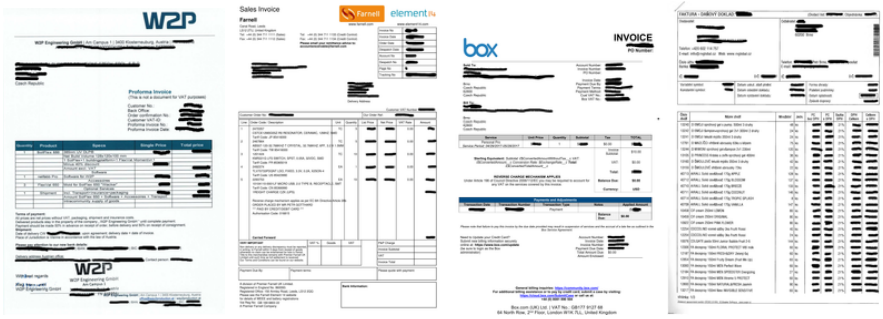

Fig. 1. Various layout formats of invoice examples

2 Classification

Workflow of the system is adapted from standard classification systems (see Figure 2). In the preprocessing step, pdf-to-image converting, image analysis (skew, quality enhancement) and languages detection are done if necessary. Then, an OCR tool is used for converting the document images into layout structures and characters: words, bounding boxes, font features and so on. After that, features are extracted to train models. In the testing phase, documents are preprocessed, features extracted and put throw trained classifier to get predicted output label.

2.1 Preprocessing

Currently, Tesseract Open Source OCR Engine (tesseract-ocr [10])[1] is among the best open source OCR engines. Since the third version, it has been supported more than 100 languages. To run tesseract-ocr from the command window:

```
tesseract imagename outputbase [-l lang] [-psm pagesegmode] [config...]
```

There are configurations for a list of languages, page segmentation mode and output format. If they are not set, then, the default are English, fully automatic page segmentation but no OSD mode, and text output relatively. Users should notice that different orders of language setting (e.g. `eng+ces` and `ces+eng`) produce different results. There are 13 selections for page segmentation mode and tesseract-ocr supports output in text, searchable PDF, hocr and tvs.

Here we run tesseract-ocr twice. The former is to detect languages used in the document. In this first run, language setting includes all possible languages of the document. For example, invoices in Czech companies usually have different versions, mostly in Czech, English, sometimes in Polish, German. Then, we base on the language distribution of words in the document to decide which languages and order of languages are used for the latter running tesseract-ocr.

[1] See `https://github.com/tesseract-ocr/tesseract/wiki/Documentation`

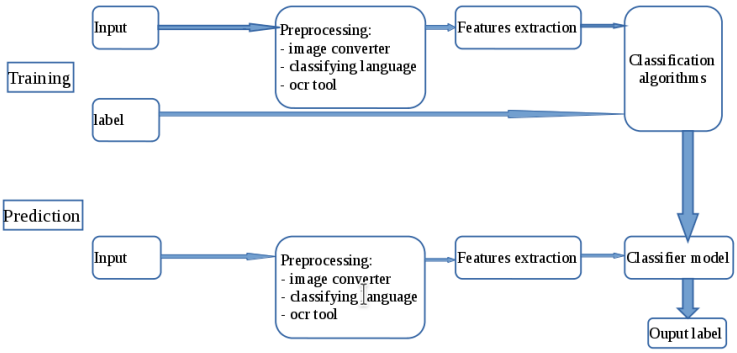

Fig. 2. Workflow of the classification system

For instance, English version invoices in Czech companies sometimes compose both English characters and Czech characters (in names, addresses). So, we set "`-l eng+ces`" for the second round running tesseract-ocr on these documents and add "`hocr`" option to get the output in hocr format.

2.2 Invoice Features

A full definition of an invoice by *The businessdictionary.com*[2] is as follows:

> "A nonnegotiable commercial instrument issued by a seller to a buyer. It identifies both the trading parties and lists, describes, and quantifies the items sold, shows the date of shipment and mode of transport, prices and discounts (if any), and delivery and payment terms.

[2] See `http://www.businessdictionary.com/definition/invoice.html`

Table 1. Data set

fold	Training set		Testing set	
	invoice	not invoice	invoice	not invoice
1	528	466	62	49
2	527	467	63	48
3	527	467	63	48
4	524	470	66	45
5	529	465	61	50
6	530	465	60	50
7	534	461	56	54
8	538	457	52	58
9	542	453	48	62
10	531	464	59	51

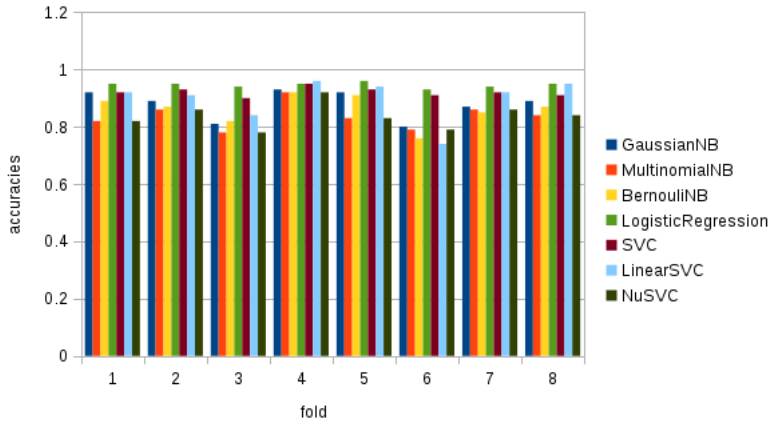

Fig. 3. Accuracies of classifiers

> In certain cases (especially when it is signed by the seller or seller's agent), an invoice serves as a demand for payment and becomes a document of title when paid in full. Types of invoice include commercial invoice, consular invoice, customs invoice, and pro-forma invoice. Also called a bill of sale or contract of sale."

Based on the definition, we can see that a typical invoice includes title "Invoice" (or "Tax Invoice", "Sale Invoice", "Proforma Invoice" and so on), a unique reference number, date of invoice, names and contact details (address, phone number, email) of providers and customers, tax payment (if relevant), description of purchased items, price, total amount, delivery and payment terms. If an invoice spreads into several pages, then there will be a page number, usually at the bottom of the page.

Table 2. Accuracies of classifiers

classifiers	1	2	3	4	5	6	7	8	9	10	mean
GaussianNB	0.92	0.89	0.81	0.93	0.92	0.87	0.93	0.80	0.87	0.89	0.89
MultinomialNB	0.82	0.86	0.78	0.92	0.83	0.76	0.90	0.79	0.86	0.84	0.83
BernouliNB	0.89	0.87	0.82	0.92	0.91	0.85	0.92	0.76	0.85	0.87	0.87
LogisticRegression	**0.95**	**0.95**	**0.94**	**0.95**	**0.96**	**0.93**	**0.95**	**0.93**	**0.94**	**0.95**	**0.95**
SVC	0.92	0.93	0.90	0.95	0.93	0.89	0.92	0.91	0.92	0.91	0.92
LinearSVC	0.92	0.91	0.84	0.96	0.94	0.85	0.93	0.74	0.92	0.95	0.92
NuSVC	0.82	0.86	0.78	0.92	0.83	0.76	0.90	0.79	0.86	0.84	0.83

Fig. 4. An example of a misclassified invoice

2.3 Classification

Assume that we have got a set of business documents including orders, invoices, emails, receipts, offer price, attachments, delivery notes and so on. The task is choosing invoices from the collection. The invoices are recognized based on their first page.

From the output of the OCR process, features are extracted. Firstly, lowercase words frequencies are counted. Then, 150 words appear most frequently on the first page of invoices are selected to be keywords. These keywords are then cleaned by removing private nouns (names, locations). After having keywords, a vector of keywords is extracted (equal 1 if the keyword is found in the document, 0 otherwise). On the other hand, the position (top) and size (width, height) of title “invoice” are extracted from the page. Here, feature "left" is not meaningful because title can be on the left, right or at the middle. If there is no such title, zero value is set for each feature. In addition, another important feature is page number. In some invoices, they keep the same general information (title, invoice number, date, supplier and customer information), just different list of items. To separate the first page from other pages, a page number is an essential pivot. If there is not any page number information, we see the page as the first page by default. Here is a piece example of feature vector from a page of an invoice:

{*”datum”: False, ”doklad”: False, ”faktura”: True, ”dič”: False, ”cena”: False, ”symbol”: True, ”dodavatel”: False, ”číslo”: True, ”spol”: True, ”variabilní”: False, ”ks”: False, ”tel”: True, ”množství”: False, ”banka”: False, ”dodací”: False, ”oddíl”: True, ”splatnosti”: True, ..., ”top”: 123, ”width”: 321, ”height”: 46, ”page”: 1*}

We try with probabilistic models such as Naive Bayes, neuron network based classifiers like Logistic Regression and Support Vector Machine.

3 Experiments

3.1 Dataset

For our experiments, 998 documents with 1505 pages are received. Out of 1505 pages, there are 1105 ones in Czech(590 first pages of invoices and 515 are not first pages of invoices), 10 in Polish and 390 in English. Polish files are pruned. Data set for English is small, so we focus on Czech ones. Invoices comes from various vendors, so the layout varies greatly (examples are in Figure 1).

3.2 Results

First of all, PDF files are converted into images. In this experiment, we use Portable Document Format to Portable Pixmap converter (pdftoppm)[3]. The default resolution of output images is 150 × 150 (dpi). Users are able to specify it to increase the quality of images. Then, these images are put through tesseract-ocr using language setting '`-l eng+ces`'. The output file is read and language distribution is counted to define the language of the document. After that phase, tesseract-ocr is used once more time with detected languages to get the words and layout format for feature extraction.

Hocr files are read into dictionaries having following keys: "text", "wordset" (including words and bounding box). Then features which are vector of keywords, top, width, height of the title "invoice" ("faktura" for Czech invoices) and page number are extracted.

To assess classifiers, we use 10-fold cross-validation. Accuracies of classifiers for each fold are listed in table 2 and depicted in Figure 3. Among classifiers, Logistic Regression scores the best with average 95.02% are recognized correctly. Support vector machines (SVC) and Linear SVC get nearly the same median

[3] See `https://freedesktop.org/wiki/Software/poppler/`

Table 3. Precision, recall and F-score of Logistic Regression model

fold	TP	FP	TN	FN	precision	recall	F-score
1	60	3	46	2	0.95	0.97	0.96
2	60	3	45	3	0.95	0.95	0.95
3	59	3	45	4	0.95	0.94	0.94
4	63	2	43	3	0.97	0.95	0.96
5	60	3	47	1	0.95	0.98	0.97
6	54	2	48	6	0.96	0.90	0.93
7	54	3	51	2	0.95	0.96	0.96
8	50	6	52	2	0.89	0.96	0.93
9	43	2	60	5	0.96	0.90	0.92
10	58	4	47	1	0.94	**0.98**	0.96
average	**58.5**	**3**	**46.5**	**2.5**	**0.95**	**0.96**	**0.95**

Table 4. Result according to feature modification

classifiers	accuracy	precision	recall	F-score
GaussNB				
all features	0.89	0.89	0.90	0.90
keywords only	0.90	0.88	**0.94**	0.92
keywords+page	**0.91**	0.89	0.94	**0.92**
title+page	0.85	**0.93**	0.78	0.85
Logistic Regression				
all feature	**0.95**	**0.95**	**0.96**	**0.95**
keywords only	0.94	0.94	0.95	0.94
keywords+page	0.94	0.94	0.95	0.95
title+page	0.84	0.91	0.79	0.84
NuSVC				
all features	0.83	0.89	0.79	0.83
keywords only	0.93	0.91	**0.96**	0.94
keywords+page	**0.93**	**0.91**	0.95	**0.94**
title+page	0.83	0.89	0.79	0.84

value but the former is more stable through folds than the latter. We should notice that samples in negative class (not a first page of an invoice) are sometimes really similar to positive samples such as page 1 and page 2 in the same invoices, or invoices and order lists.

Having a close look at Logistic Regression model's results, most of files are correctly recognized and there are very few false negative (average 2.5(4%)) and false positive (average 3(6%)). Detail data is in table 3.

Errors are partly because of OCR errors. For example, in the invoice in Figure 4, the title "FAKTURA" (invoice) is wrong recognized as "FAgTURA". Therefore, title position features are set zeros. In this situation, apart from Logistic Regression and Linear SVC, all other classifiers classify it as not a first page of an invoice. Logistic Regression model predicts right 8 out of 10 times whereas Linear SVC scores 10/10.

Surprisingly, when we remove title position features and only keep keywords or keywords and a page number, all models, except Logistic Regression, have improvements in accuracy, recall and F-score, particularly recall because of looser constrains. Examples of changes on average measurements can be seen in table 4.

4 Conclusion and Future Work

In this paper, we have built a classification system to recognize the first page of invoices from scanned documents. Based on used features, it can be adapted for other languages. This work mainly uses words, the smallest unit in document layout, to extract features. In subsequent work, we will construct

blocks of information based on available features (geometric and textual features of word, line), and use Natural Language Processing tools such as named entity recognition to extract semantic meaning of blocks. This will provide important features for classification as well as information extraction from scanned invoices.

5 Acknowledgements

This work has been partly supported by Konica Minolta Business Solution Czech within the OCR Miner project and by the Masaryk University project MUNI/33/55939/2017.

References

1. Ronen Feldman and James Sanger: The Text Mining Handbook: Advanced approaches in analyzing unstructured data. Cambridge University Press, New York, 2007.
2. Stefan Klink, Andreas Dengel, Thomas Kieninger: Document Structure Analysis Based on Layout and Textual Features. In: Proc. of International Workshop on Document Analysis Systems, DAS2000, 2000.
3. T.A Bayer and H.U.Mogg-Schneider: A generic system for processing invoices. IEEE, 1997.
4. MARINAI, Simone: Introduction to document analysis and recognition. Machine learning in document analysis and recognition, 2008, 1–20.
5. Fabrizio Sebastiani. Machine learning in automated text categorization. ACM Computing surveys, vol. 34, No. 1, March 2002, pp.1–47.
6. S.L. Ting, W.H. Ip, Albert H.C. Tsang. Is Naïve Bayes a Good Classifier for Document Classification? International Journal of Software Engineering and Its Applications Vol. 5, No. 3, July, 2011.
7. AGGARWAL, Charu C. and ZHAI, ChengXiang. A survey of text classification algorithms. Mining text data, 2012, 163–222.
8. HAMZA, Hatem, BELAÏD, Yolande and BELAÏD, Abdel. Case-based reasoning for invoice analysis and recognition. In: ICCBR. p. 404–418. 2007.
9. SCHULZ, Frederick, et al. Seizing the treasure: Transferring knowledge in invoice analysis. In: Document Analysis and Recognition, ICDAR'09, 10th International Conference on. IEEE, p. 848–852, 2009.
10. Smith, Ray: An overview of the Tesseract OCR engine, the Ninth International Conference on Document Analysis and Recognition, ICDAR 2007, vol. 2, pp. 629–633, IEEE, 2007.