Recurrent Neural Network Approach for Table Field Extraction in Business Documents

Clément Sage, Alexandre Aussem, Haytham Elghazel, Véronique Eglin and Jérémy Espinas

Univ Lyon, CNRS, LIRIS and Esker

Objective

Extract tabular information from business documents under the following conditions:

- Predefined and fixed set of information types for a given document class
- Positioning and textual representation of the information, i.e. templates, are unconstrained

Context

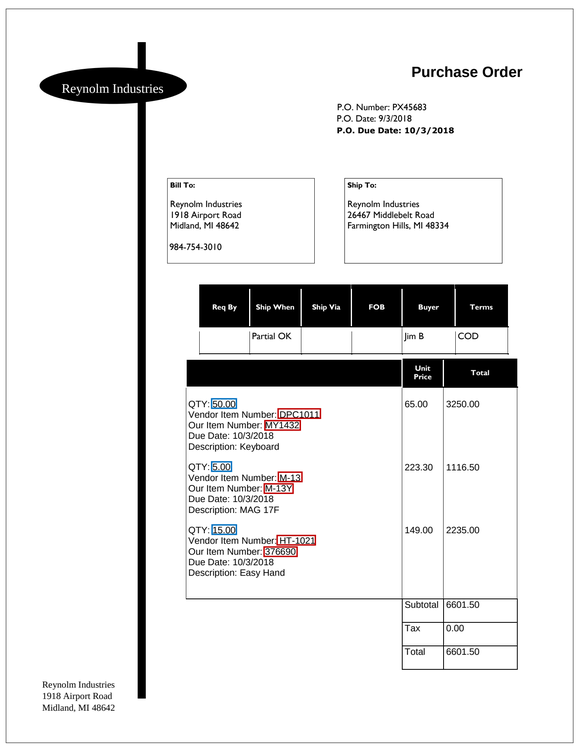


Figure 1. Illustration of our experimental task: Extract the ID numbers (red) and quantities (blue) of products contained in purchase orders.

If performed manually, extracting information from their incoming documents is a daunting task for companies. In our work, we aim at automating this step. Particularly, we focus on the extraction of two fields contained in tables of ordered products (Figure 1).

Related works

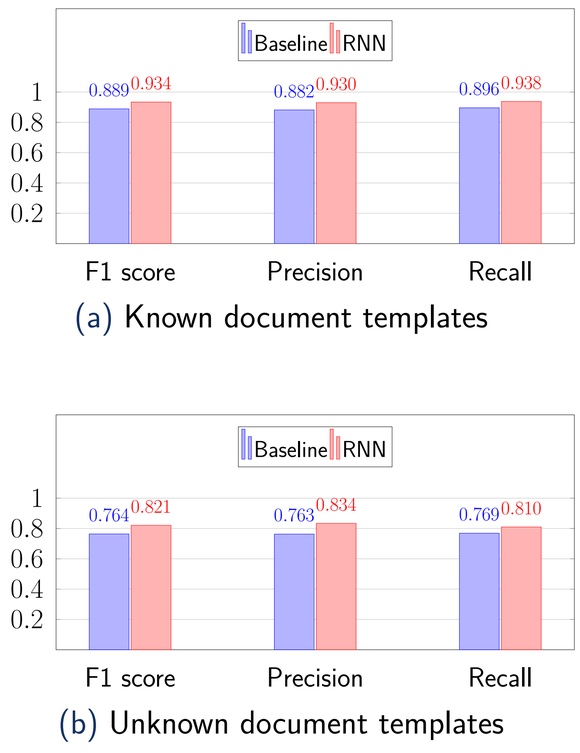
- Template incremental learning methods [1, 2] that detect templates and apply template specific extraction models
- Template agnostic methods:
  - Based on domain specific knowledge encoded in hand-crafted features [3]
  - Feature-inferring Recurrent Neural Networks (RNN) which are the state-of-the-art to extract non tabular data in invoices [4]

Our contribution: Prove that a template agnostic RNN approach is also relevant for extracting the less tackled table fields

Approach

See Figure 3 for an overview of our approach:

1. Apply an external OCR engine on the document if needed
2. Attribute a vocabulary index to each word based on its normalized textual value
3. Concatenate the dense learnable embedding associated to its index with spatial and case features
4. BLSTM layers iterate on the 1D Z-ordered sequence of words
5. A $k+1$ unit softmax layer predicts the word classes, k being the number of field types to extract



(a) Known document templates

(b) Unknown document templates

Figure 2. Extraction metrics evaluated on the test set and micro averaged over the ID Number and Quantity fields.

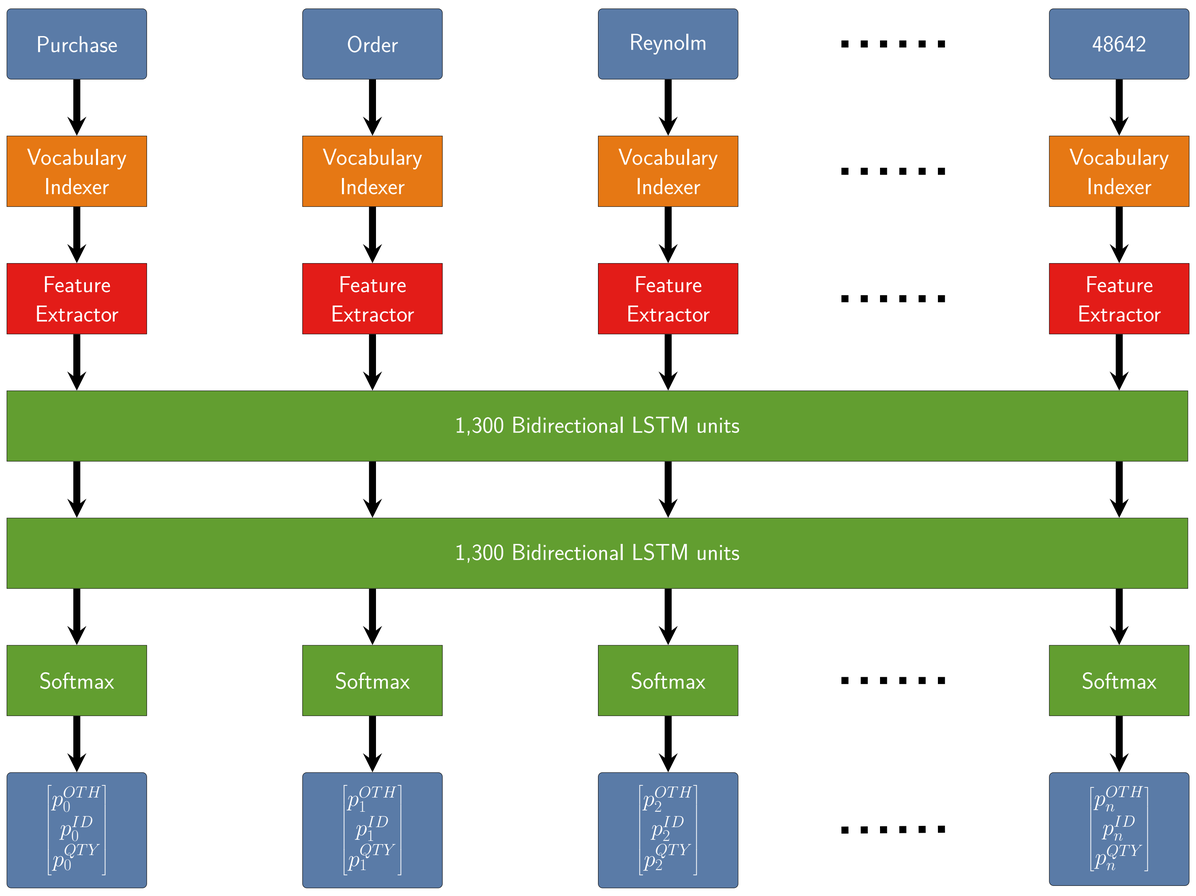


Figure 3. Our approach for extracting the ID numbers and quantities (resp. abbreviated ID and QTY) of the purchase order given in Figure 1. OTH class gathers information types that we do not want to extract.

Experiments

Our approach is evaluated on a private dataset of 28,570 English purchase orders representing 2,818 templates.

We split this dataset according to two ways such that the training and test sets have either:

a. common templates

b. different templates

We compare our method with its feedforward network equivalent. This baseline is enhanced with local context knowledge encoded in its input. To do so, the feature vectors of each word are concatenated with the vectors of 4 of their spatially closest words.

Results

As shown by Figure 2, the RNN substantially surpasses the feedforward baseline in terms of micro precision, recall and F1 score, thus demonstrating the usefulness of recurrent connections for extracting table fields.

As expected, extraction performances are higher for known templates (Figure 2a) than for unknown templates (Figure 2b), with respective micro F1 scores of 0.934 and 0.821 for the RNN model. However, the difference is rather small compared to performance gaps observed for template incremental learning methods [2].

Conclusions

We experimentally showed that a word level RNN combining textual and spatial features is effective for retrieving table fields, even for document templates not seen during model training.

Perspectives

- Provide a more granular and generic parsing of the text with a character level RNN generating textual components of the word feature vectors
- Tackle recognition of structured tabular entities to extract products from business documents by exploring encoder-decoder RNN architectures augmented with attention mechanisms [5]
- Assess our extraction models when confronted with a multilingual dataset

References

[1] V. P. d'Andecy, E. Hartmann, and M. Rusiñol, "Field extraction by hybrid incremental and a-priori structural templates," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 2018, pp. 251–256.

[2] D. Esser, D. Schuster, K. Muthmann, and A. Schill, "Few-exemplar information extraction for business documents." in *ICEIS (1)*, 2014, pp. 293–298.

[3] F. Deckert, B. Seidler, M. Ebbecke, and M. Gillmann, "Table content understanding in smartfix," in *2011 International Conference on Document Analysis and Recognition*. IEEE, 2011, pp. 488–492.

[4] R. B. Palm, O. Winther, and F. Laws, "Cloudscan-a configuration-free invoice analysis system using recurrent neural networks," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017, pp. 406–413.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

Contact Information

- clement.sage@liris.cnrs.fr
- +33 (0)6 98 57 61 81