RECURRENT NEURAL NETWORK APPROACH FOR TABLE FIELD EXTRACTION IN BUSINESS DOCUMENTS

PhD Student: Clément Sage

Supervisors: Alexandre Aussem, Haytham Elghazel, Véronique Eglin and Jérémy Espinas

Université Claude Bernard Lyon 1

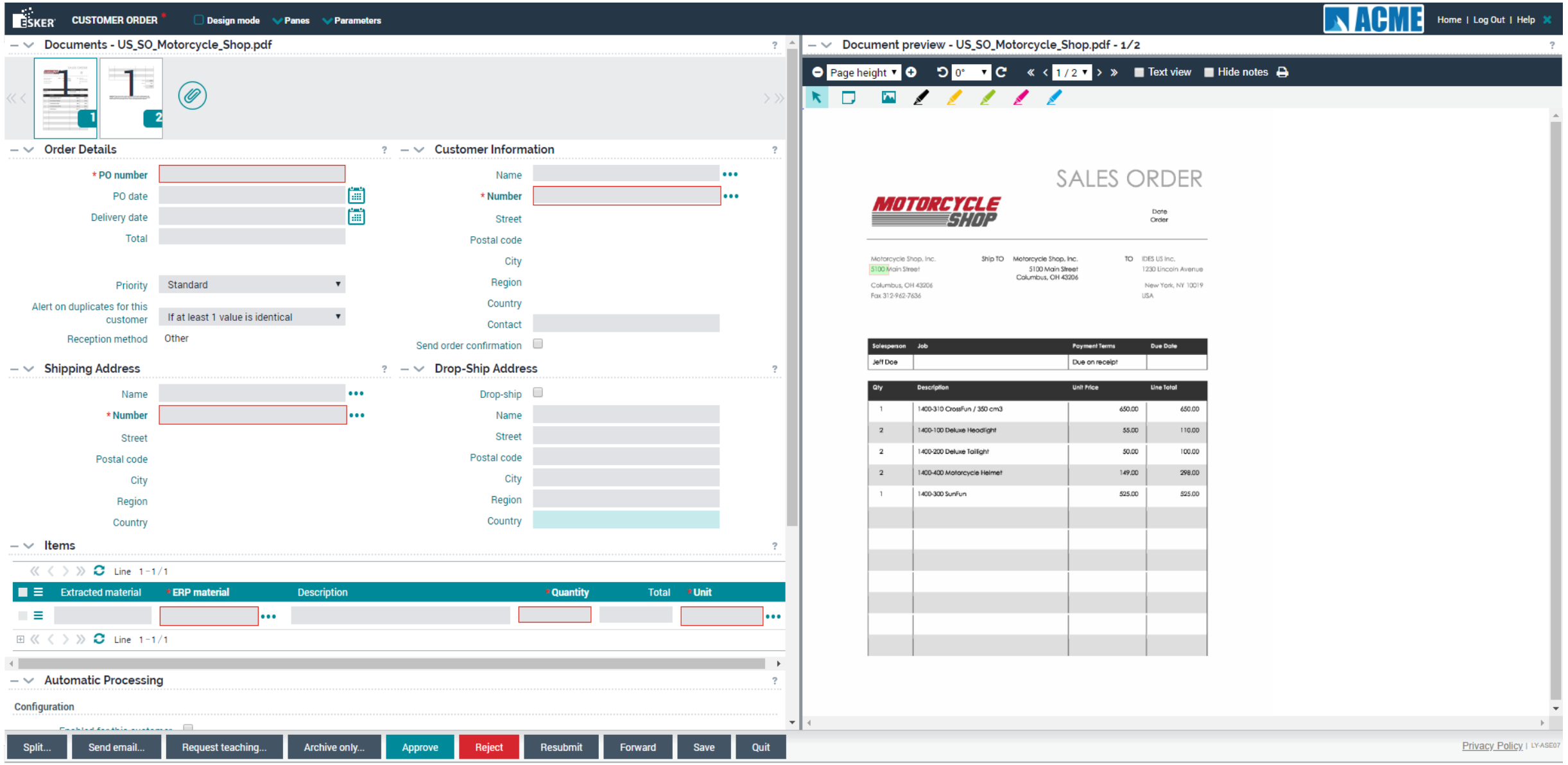
LIRIS

ESKER

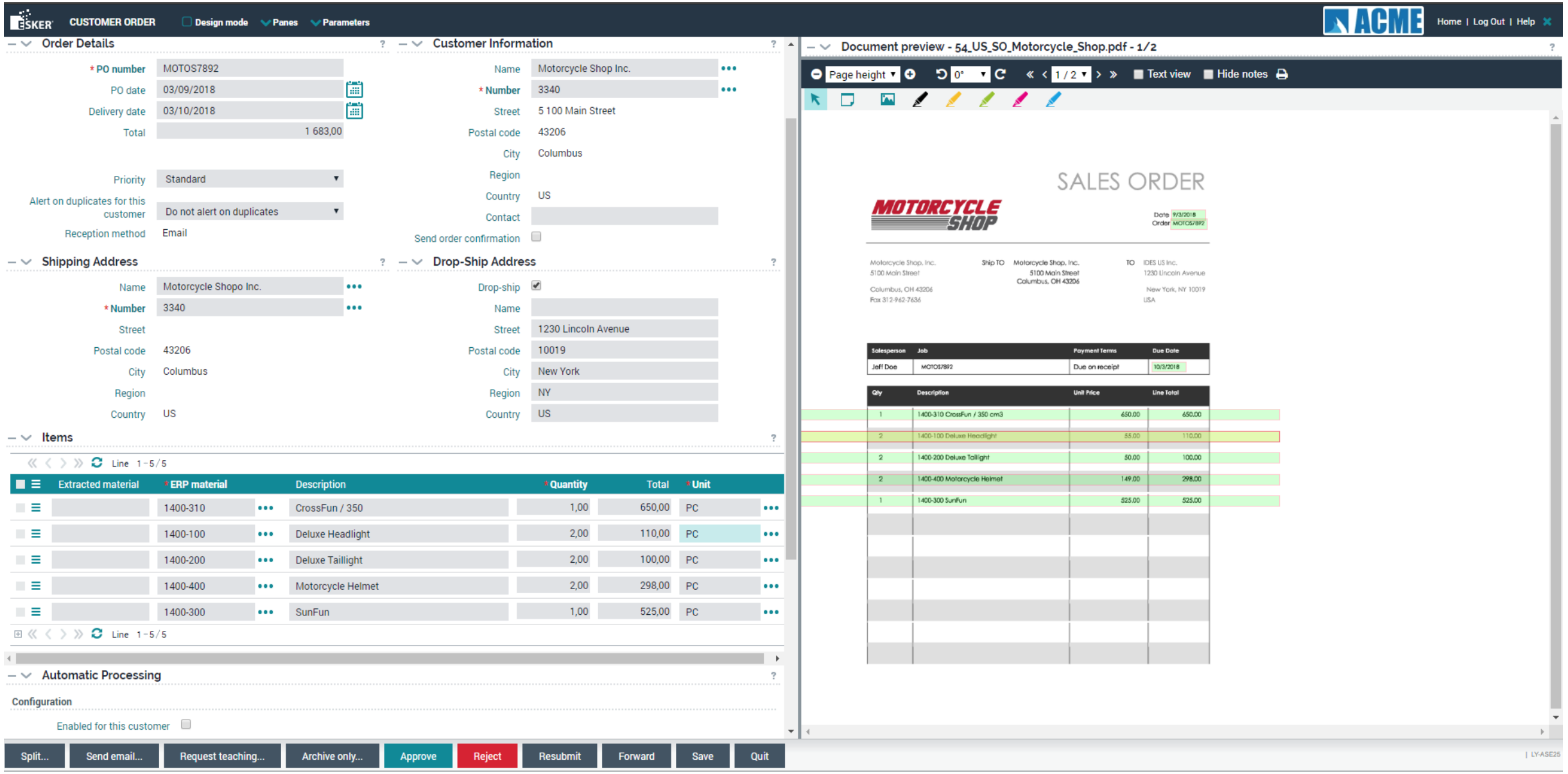
OUTLINE OF THE PRESENTATION

- Information extraction in business documents
- State of the art review
- Presentation of our method
- Dataset & experiments
- Results
- Conclusions & perspectives

INFORMATION EXTRACTION (IE) IN BUSINESS DOCUMENTS



INFORMATION EXTRACTION (IE) IN BUSINESS DOCUMENTS



INFORMATION EXTRACTION CONSTRAINTS IMPOSED BY INDUSTRIAL CONTEXT

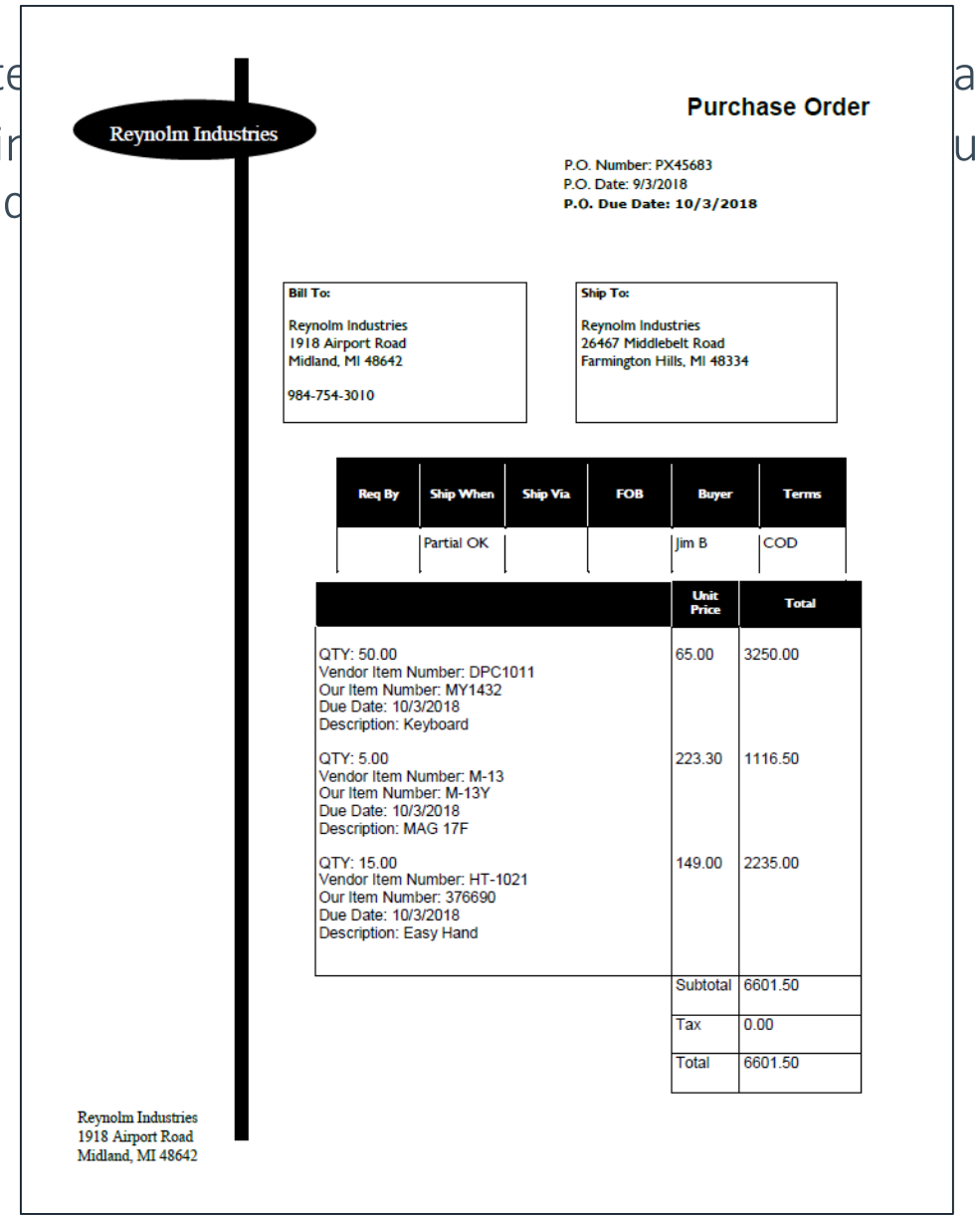
- Able to operate on both scanned and born-digital documents
- Should be generic enough to be easily adapted for any language, document class or information to extract
- Should accurately extract information even for a document layout (also called template) which has not already processed by the system
  - Structural Template Incremental Learning approaches such as in [1, 2, 3, 4, 5] cannot be employed
  - Should develop a template free method

STATE OF THE ART REVIEW OF TEMPLATE FREE IE APPROACHES

- Lot of recent works on the detection and physical structure recognition of tables in documents [6, 7, 8]
  - However, table structure in business documents is frequently ambiguous and may not provide enough knowledge for further field extraction

STATE OF THE ART REVIEW OF TEMPLATE FREE IE APPROACHES

- Lot of recent works on the dete ables in documents [6, 7, 8]
  - › However, table structure ir us and may not provide enough knowledge for further field



Reynolm Industries

Purchase Order

P.O. Number: PX45683
P.O. Date: 9/3/2018
P.O. Due Date: 10/3/2018

Bill To:

Reynolm Industries
1918 Airport Road
Midland, MI 48642

984-754-3010

Ship To:

Reynolm Industries
26467 Middlebelt Road
Farmington Hills, MI 48334

Req By	Ship When	Ship Via	FOB	Buyer	Terms
	Partial OK			Jim B	COD

	Unit Price	Total
QTY: 50.00 Vendor Item Number: DPC1011 Our Item Number: MY1432 Due Date: 10/3/2018 Description: Keyboard	65.00	3250.00
QTY: 5.00 Vendor Item Number: M-13 Our Item Number: M-13Y Due Date: 10/3/2018 Description: MAG 17F	223.30	1116.50
QTY: 15.00 Vendor Item Number: HT-1021 Our Item Number: 376690 Due Date: 10/3/2018 Description: Easy Hand	149.00	2235.00
	Subtotal	6601.50
	Tax	0.00
	Total	6601.50

Reynolm Industries
1918 Airport Road
Midland, MI 48642

STATE OF THE ART REVIEW OF TEMPLATE FREE IE APPROACHES

- Several works on the detection and physical structure recognition of tables in documents [6, 7, 8]:
  - However, table structure in business documents is frequently ambiguous and may not provide enough knowledge for further field extraction

- Methods [9, 10, 11] that rely on significant domain specific knowledge and a number of handcrafted features or rules:
  - Showing troubles when tackling highly unstructured fields in documents with strong layout variability

STATE OF THE ART REVIEW OF TEMPLATE FREE IE APPROACHES

- Several works on the detection and physical structure recognition of tables in documents [6, 7, 8]:
  - However, table structure in business documents is frequently ambiguous and may not provide enough knowledge for further field extraction

- Methods [9, 10, 11] that rely on significant domain specific knowledge and a number of handcrafted features or rules:
  - Showing troubles when tackling highly unstructured fields in documents with strong layout variability

- IE is closely related with the Named Entity Recognition (NER) task which is actively studied [13, 14, 15]:
  - However, IE in business documents presents some specificities: word alignment and spacing, reading order, grammar

STATE OF THE ART REVIEW OF TEMPLATE FREE IE APPROACHES

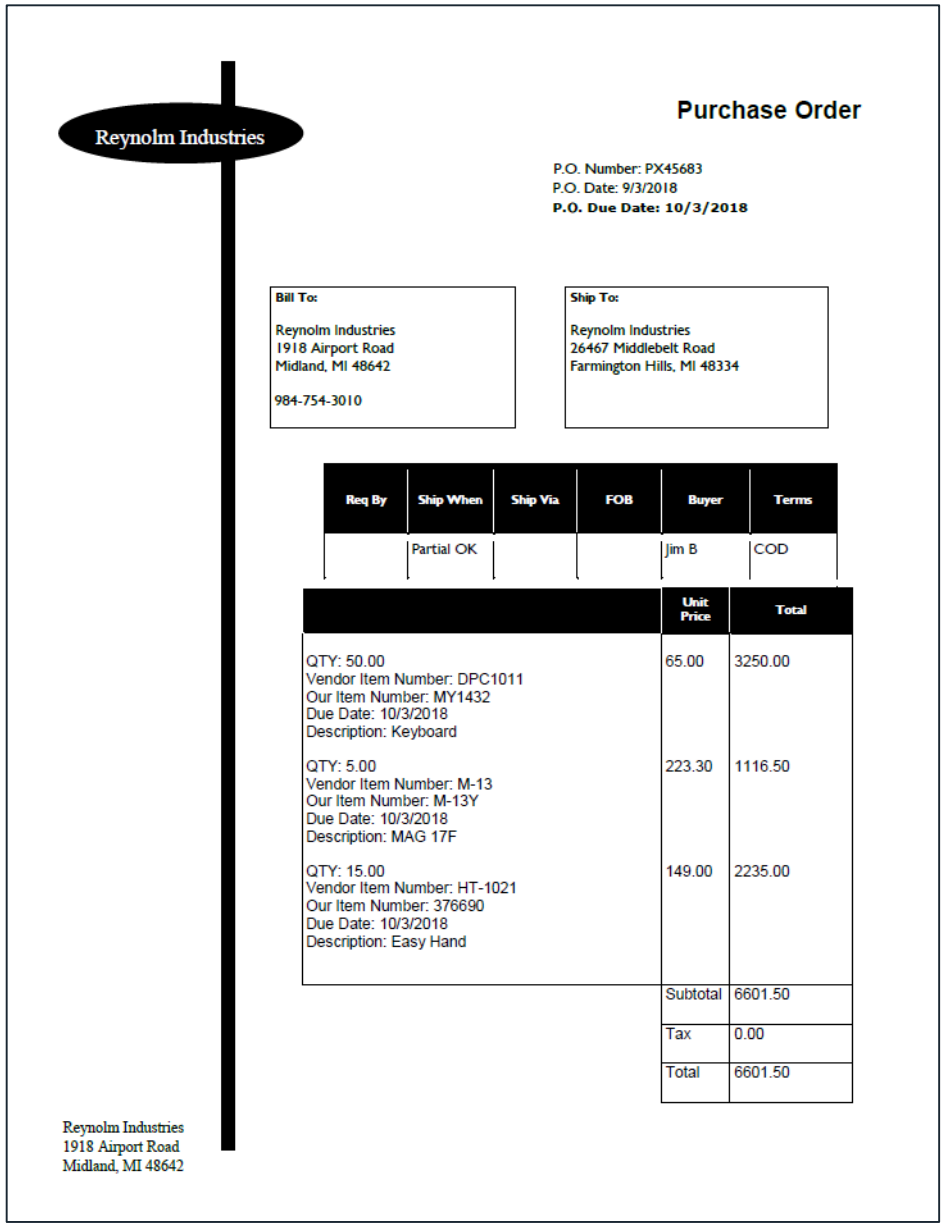
- Several works on the detection and physical structure recognition of tables in documents [6, 7, 8]:
  - However, table structure in business documents is frequently ambiguous and may not provide enough knowledge for further field extraction

- Methods [9, 10, 11] that rely on significant domain specific knowledge and a number of handcrafted features or rules:
  - Showing troubles when tackling highly unstructured fields in documents with strong layout variability

- IE is closely related with the Named Entity Recognition (NER) task which is actively studied [13, 14, 15]:
  - However, IE in business documents presents some specificities: word alignment and spacing, reading order, grammar

- Sequence labelling approach with recurrent neural network (RNN) [12] for extracting non tabular data in invoices:
  - Applying this approach to tabular data, which are more challenging to retrieve ?

TABLE FIELD EXTRACTION OBJECTIVE



Reynolm Industries

Purchase Order

P.O. Number: PX45683
P.O. Date: 9/3/2018
P.O. Due Date: 10/3/2018

Bill To:

Reynolm Industries
1918 Airport Road
Midland, MI 48642

984-754-3010

Ship To:

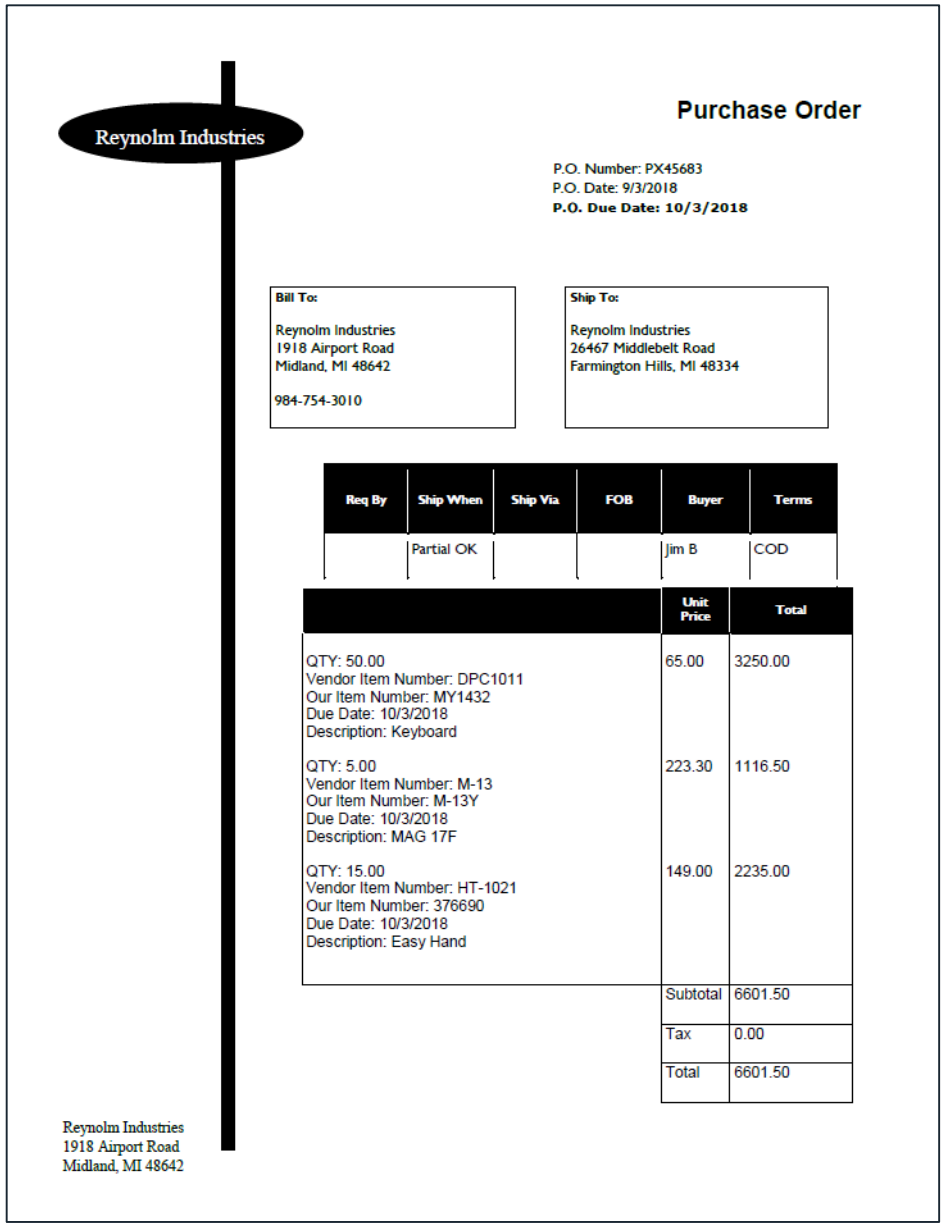
Reynolm Industries
26467 Middlebelt Road
Farmington Hills, MI 48334

Req By	Ship When	Ship Via	FOB	Buyer	Terms
	Partial OK			Jim B	COD

	Unit Price	Total
QTY: 50.00 Vendor Item Number: DPC1011 Our Item Number: MY1432 Due Date: 10/3/2018 Description: Keyboard	65.00	3250.00
QTY: 5.00 Vendor Item Number: M-13 Our Item Number: M-13Y Due Date: 10/3/2018 Description: MAG 17F	223.30	1116.50
QTY: 15.00 Vendor Item Number: HT-1021 Our Item Number: 376690 Due Date: 10/3/2018 Description: Easy Hand	149.00	2235.00
	Subtotal	6601.50
	Tax	0.00
	Total	6601.50

Reynolm Industries
1918 Airport Road
Midland, MI 48642

TABLE FIELD EXTRACTION OBJECTIVE



Reynolm Industries

Purchase Order

P.O. Number: PX45683
P.O. Date: 9/3/2018
P.O. Due Date: 10/3/2018

Bill To:

Reynolm Industries
1918 Airport Road
Midland, MI 48642

984-754-3010

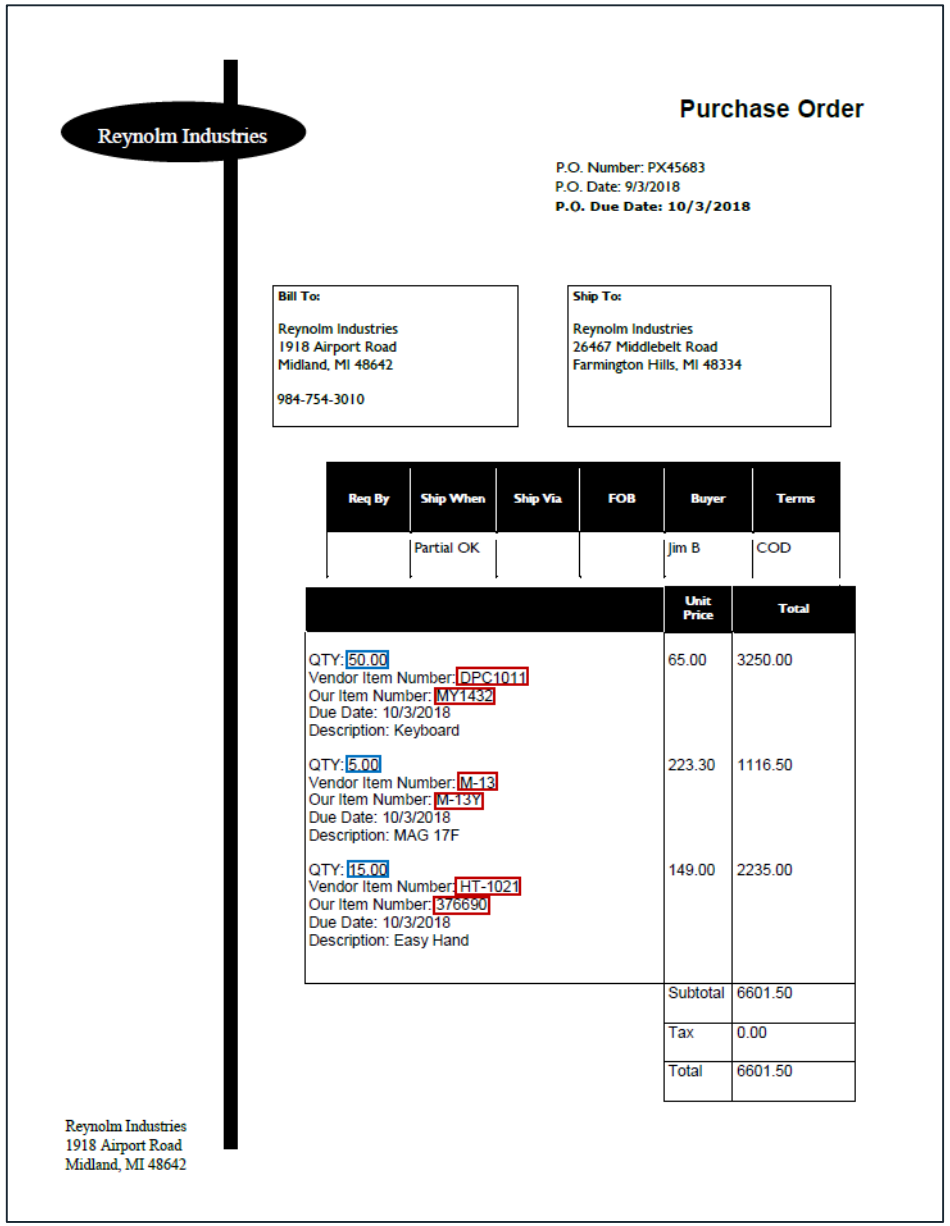
Ship To:

Reynolm Industries
26467 Middlebelt Road
Farmington Hills, MI 48334

Req By	Ship When	Ship Via	FOB	Buyer	Terms
	Partial OK			Jim B	COD

	Unit Price	Total
QTY: 50.00 Vendor Item Number: DPC1011 Our Item Number: MY1432 Due Date: 10/3/2018 Description: Keyboard	65.00	3250.00
QTY: 5.00 Vendor Item Number: M-13 Our Item Number: M-13Y Due Date: 10/3/2018 Description: MAG 17F	223.30	1116.50
QTY: 15.00 Vendor Item Number: HT-1021 Our Item Number: 376690 Due Date: 10/3/2018 Description: Easy Hand	149.00	2235.00
	Subtotal	6601.50
	Tax	0.00
	Total	6601.50

Reynolm Industries
1918 Airport Road
Midland, MI 48642



Reynolm Industries

Purchase Order

P.O. Number: PX45683
P.O. Date: 9/3/2018
P.O. Due Date: 10/3/2018

Bill To:

Reynolm Industries
1918 Airport Road
Midland, MI 48642

984-754-3010

Ship To:

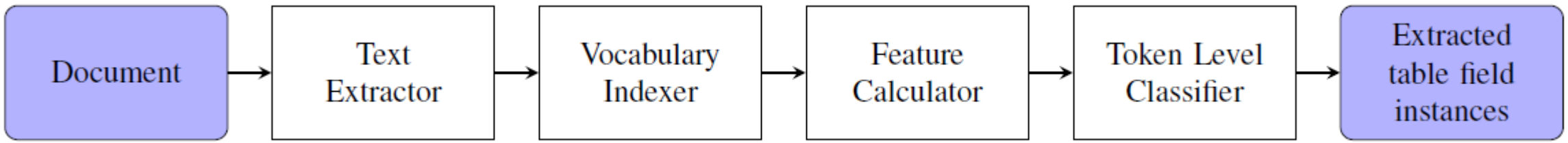
Reynolm Industries
26467 Middlebelt Road
Farmington Hills, MI 48334

Req By	Ship When	Ship Via	FOB	Buyer	Terms
	Partial OK			Jim B	COD

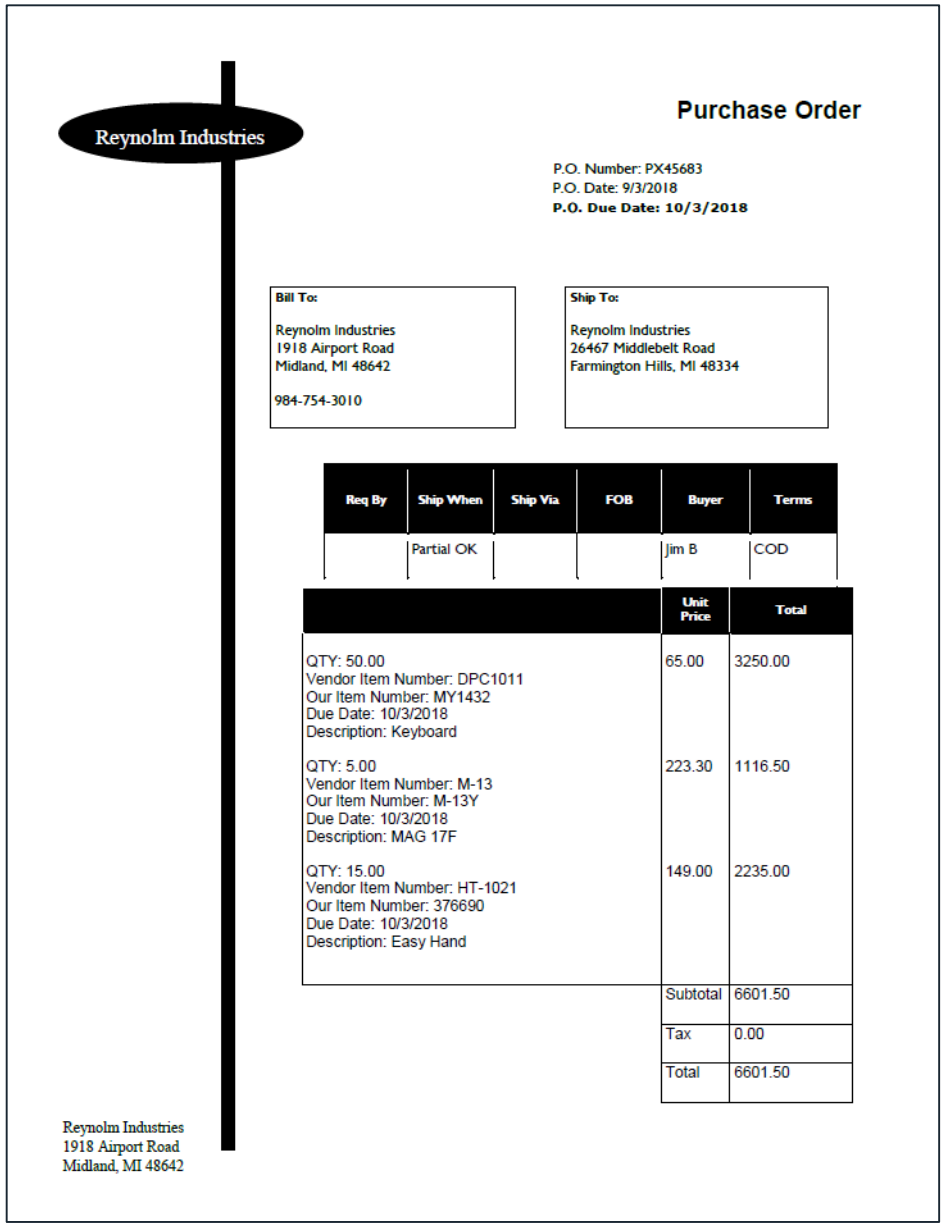
	Unit Price	Total
QTY: 50.00 Vendor Item Number: DPC1011 Our Item Number: MY1432 Due Date: 10/3/2018 Description: Keyboard	65.00	3250.00
QTY: 5.00 Vendor Item Number: M-13 Our Item Number: M-13Y Due Date: 10/3/2018 Description: MAG 17F	223.30	1116.50
QTY: 15.00 Vendor Item Number: HT-1021 Our Item Number: 376690 Due Date: 10/3/2018 Description: Easy Hand	149.00	2235.00
	Subtotal	6601.50
	Tax	0.00
	Total	6601.50

Reynolm Industries
1918 Airport Road
Midland, MI 48642

OVERVIEW OF OUR METHOD



STEP #1: TEXT EXTRACTOR



Reynolm Industries

Purchase Order

P.O. Number: PX45683
P.O. Date: 9/3/2018
P.O. Due Date: 10/3/2018

Bill To:

Reynolm Industries
1918 Airport Road
Midland, MI 48642

984-754-3010

Ship To:

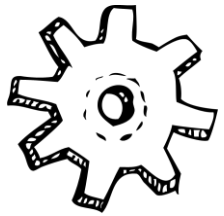
Reynolm Industries
26467 Middlebelt Road
Farmington Hills, MI 48334

Req By	Ship When	Ship Via	FOB	Buyer	Terms
	Partial OK			Jim B	COD

	Unit Price	Total
QTY: 50.00 Vendor Item Number: DPC1011 Our Item Number: MY1432 Due Date: 10/3/2018 Description: Keyboard	65.00	3250.00
QTY: 5.00 Vendor Item Number: M-13 Our Item Number: M-13Y Due Date: 10/3/2018 Description: MAG 17F	223.30	1116.50
QTY: 15.00 Vendor Item Number: HT-1021 Our Item Number: 376690 Due Date: 10/3/2018 Description: Easy Hand	149.00	2235.00
	Subtotal	6601.50
	Tax	0.00
	Total	6601.50

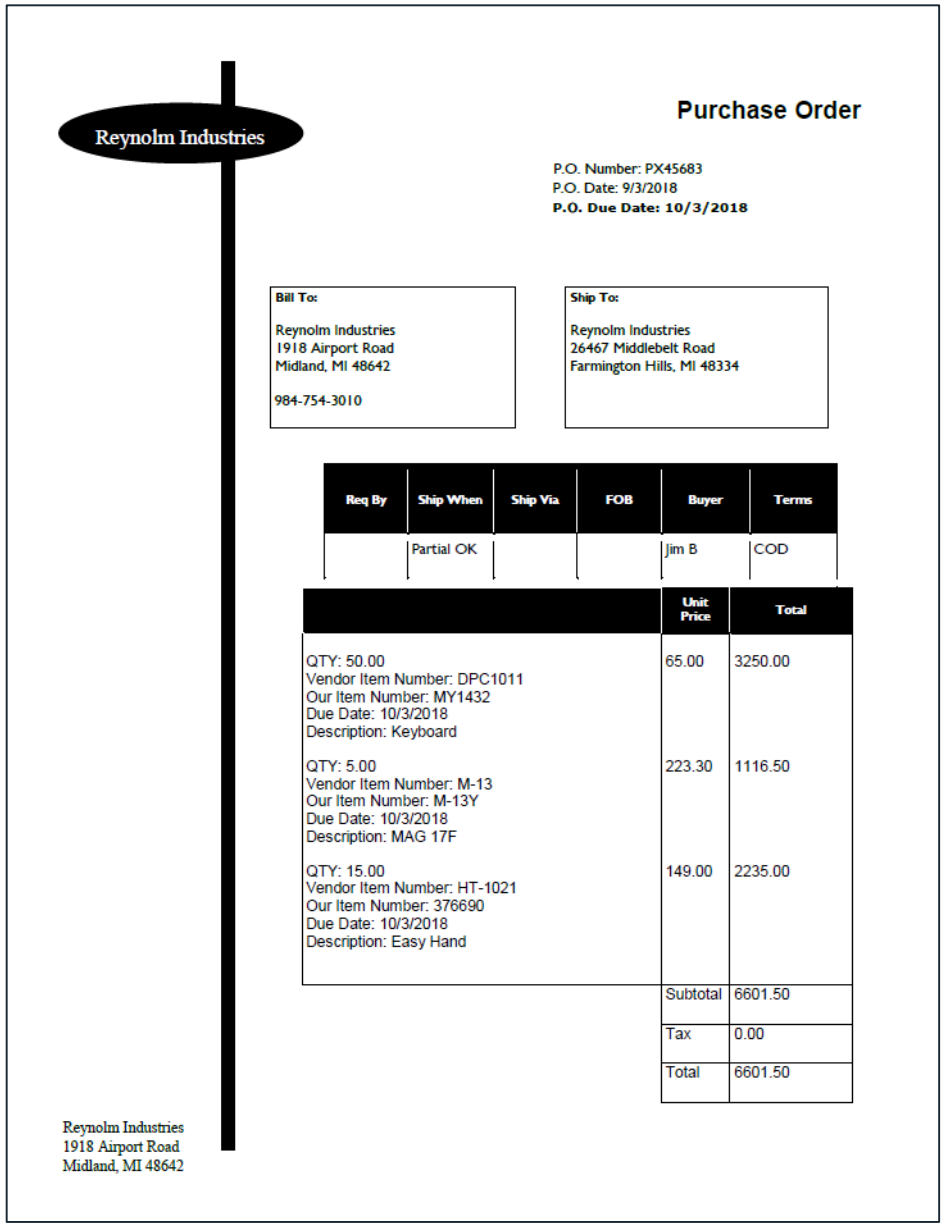
Reynolm Industries
1918 Airport Road
Midland, MI 48642

External OCR



engine

STEP #1: TEXT EXTRACTOR



Reynolm Industries

Purchase Order

P.O. Number: PX45683
P.O. Date: 9/3/2018
P.O. Due Date: 10/3/2018

Bill To:

Reynolm Industries
1918 Airport Road
Midland, MI 48642

984-754-3010

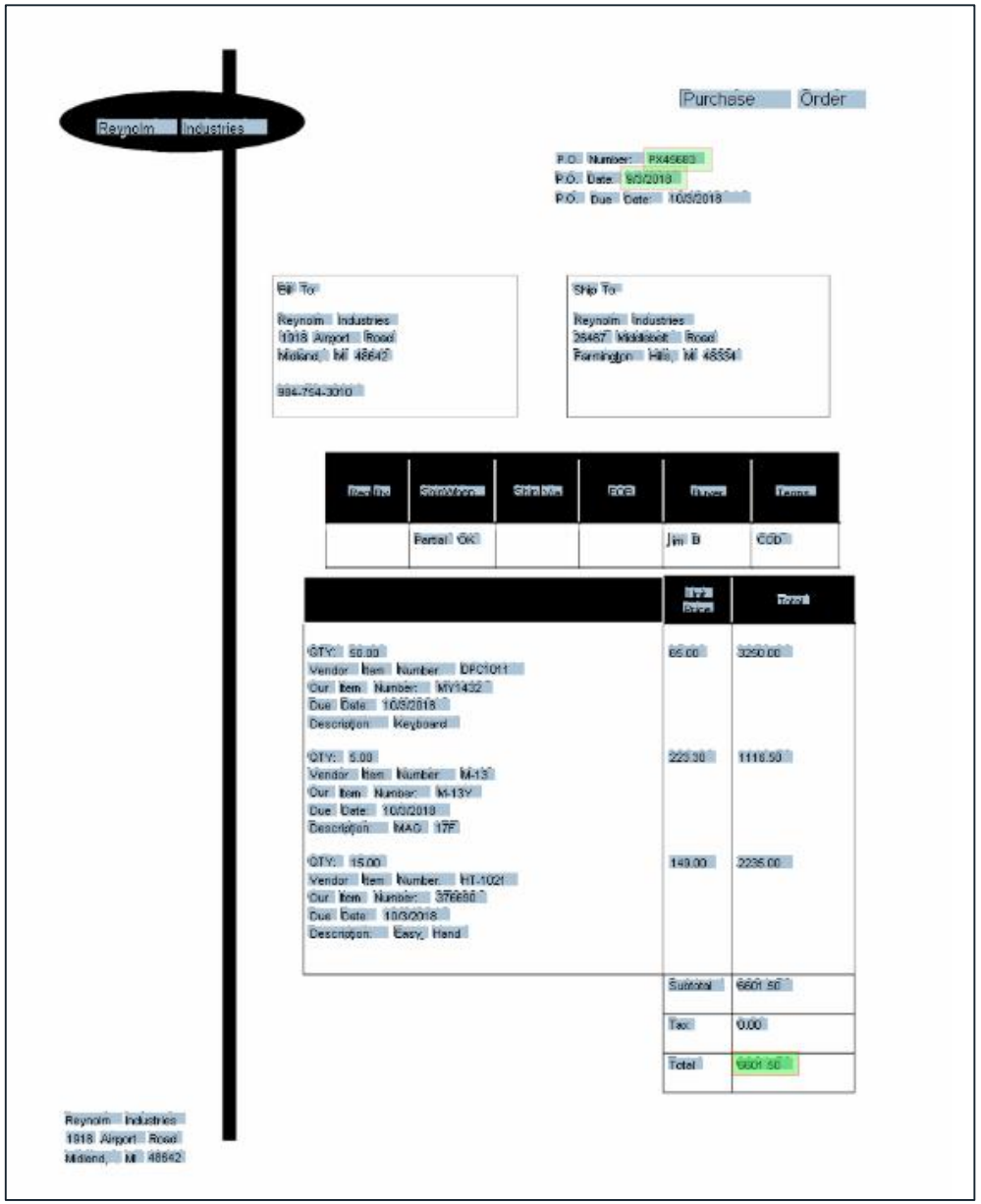
Ship To:

Reynolm Industries
26467 Middlebelt Road
Farmington Hills, MI 48334

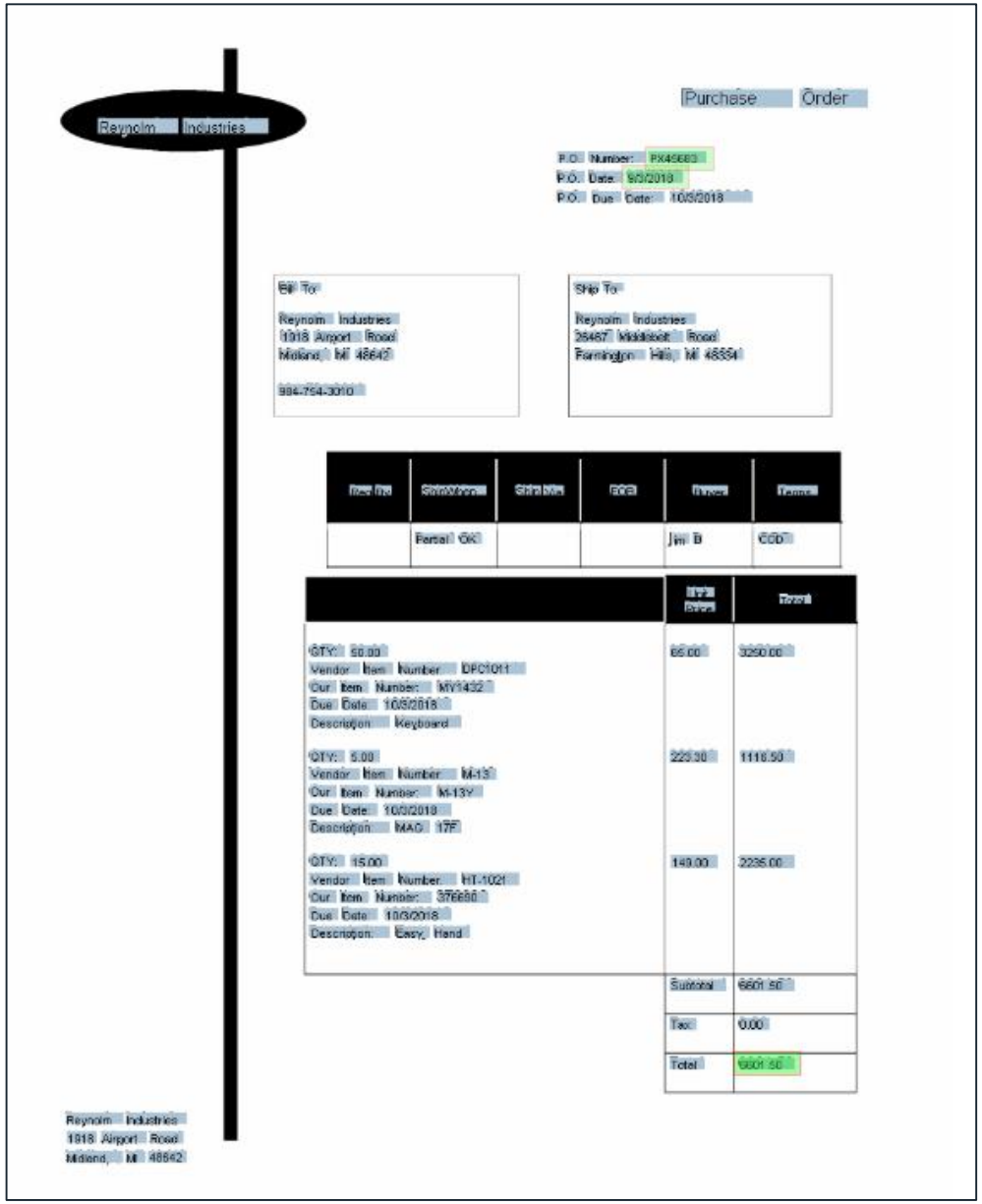
Req By	Ship When	Ship Via	FOB	Buyer	Terms
	Partial OK			Jim B	COD

	Unit Price	Total
QTY: 50.00 Vendor Item Number: DPC1011 Our Item Number: MY1432 Due Date: 10/3/2018 Description: Keyboard	65.00	3250.00
QTY: 5.00 Vendor Item Number: M-13 Our Item Number: M-13Y Due Date: 10/3/2018 Description: MAG 17F	223.30	1116.50
QTY: 15.00 Vendor Item Number: HT-1021 Our Item Number: 376690 Due Date: 10/3/2018 Description: Easy Hand	149.00	2235.00
	Subtotal	6601.50
	Tax	0.00
	Total	6601.50

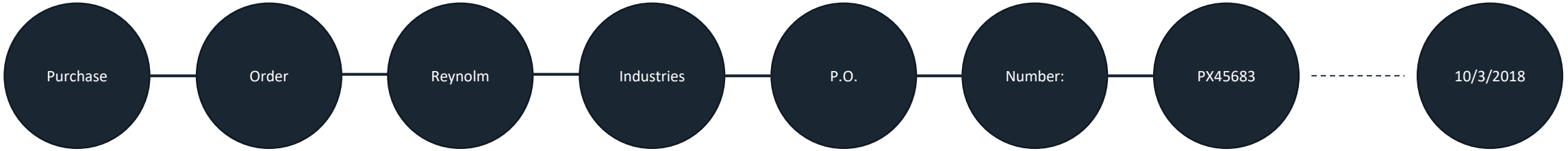
Reynolm Industries
1918 Airport Road
Midland, MI 48642



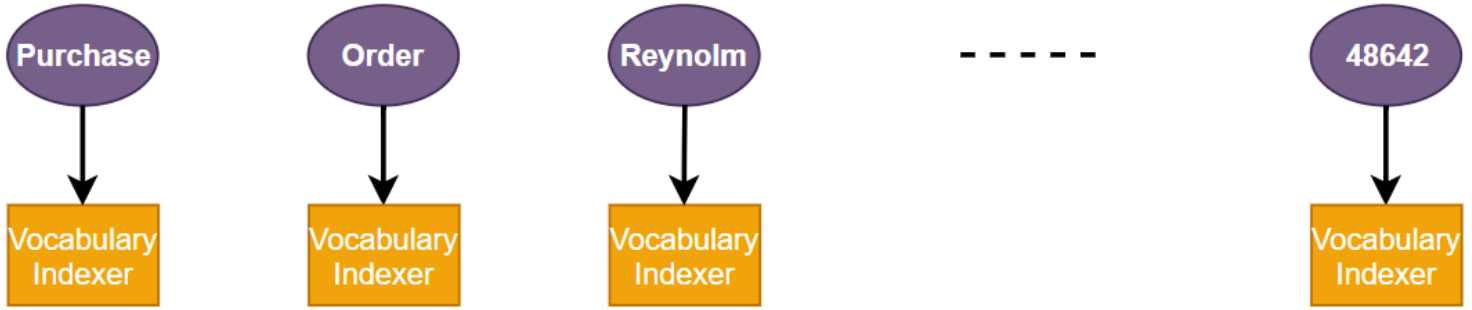
STEP #1: TEXT EXTRACTOR, ORDERING OF THE TOKENS IN Z ORDER



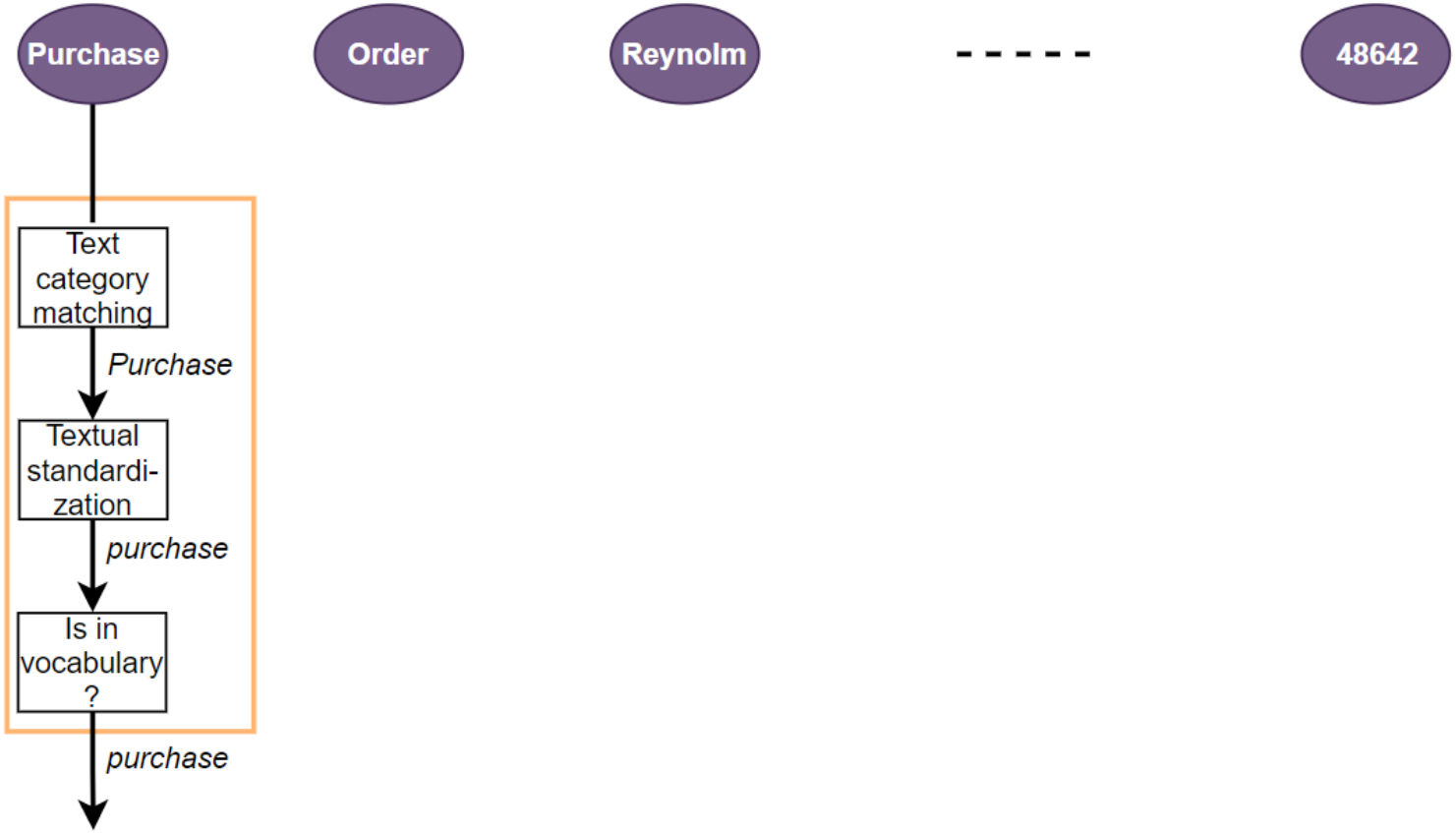
STEP #1: TEXT EXTRACTOR, ORDERING OF THE TOKENS IN Z ORDER



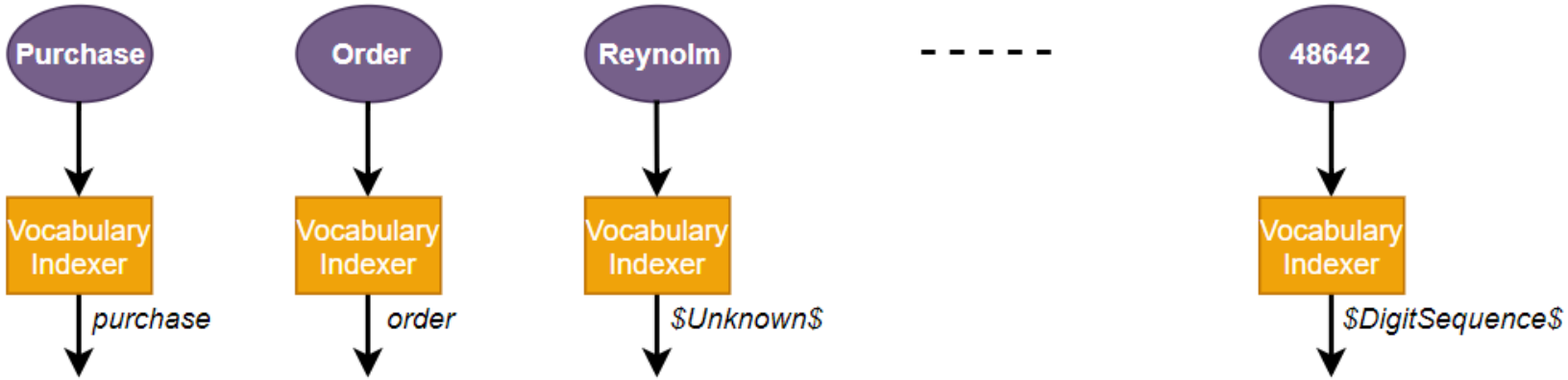
STEP #2: VOCABULARY INDEXER



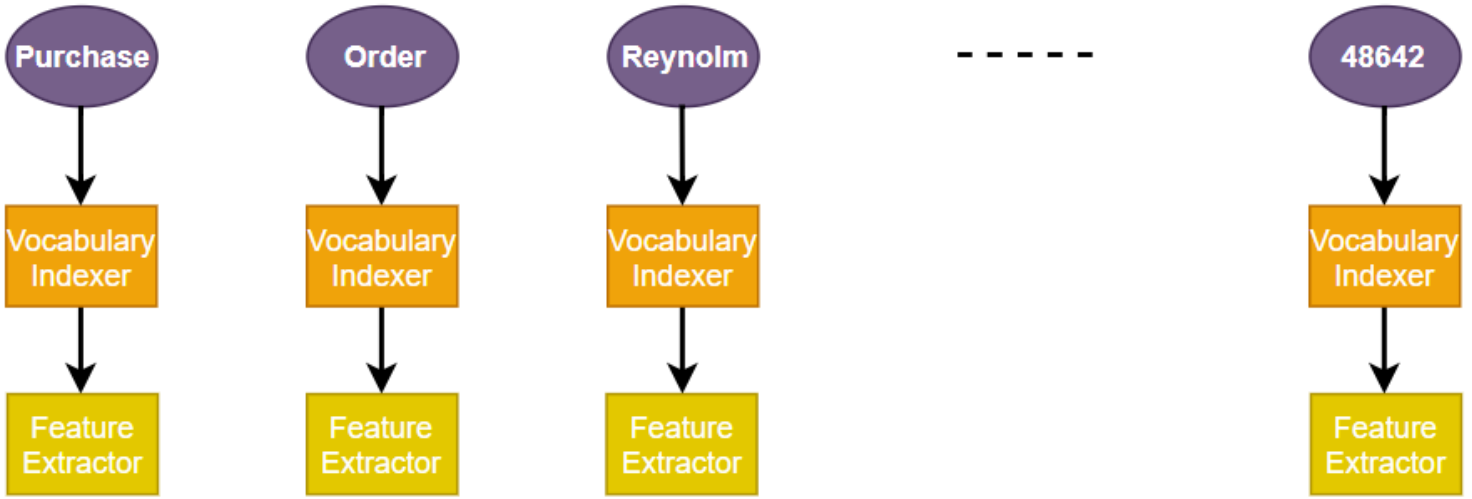
STEP #2: VOCABULARY INDEXER



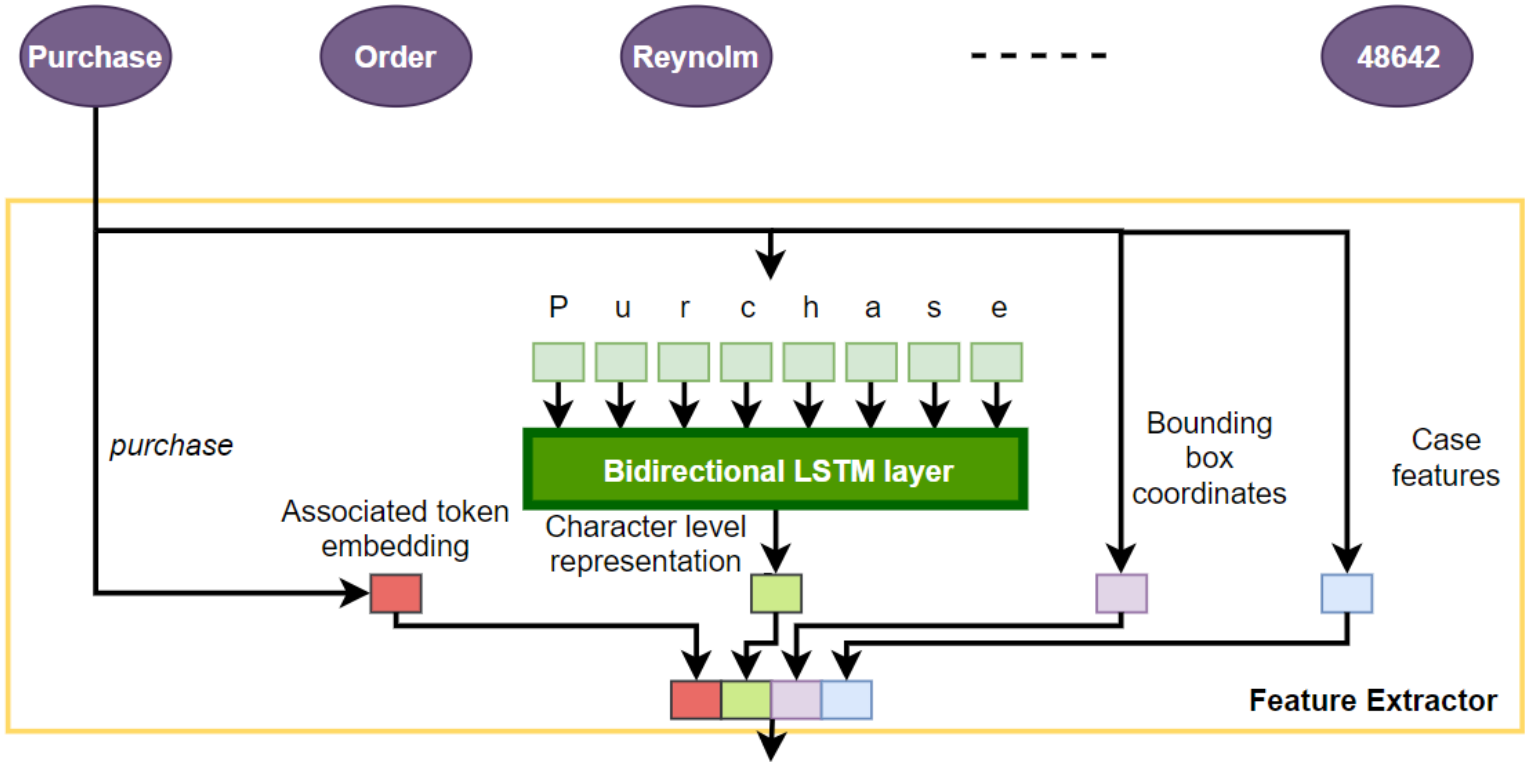
STEP #2: VOCABULARY INDEXER



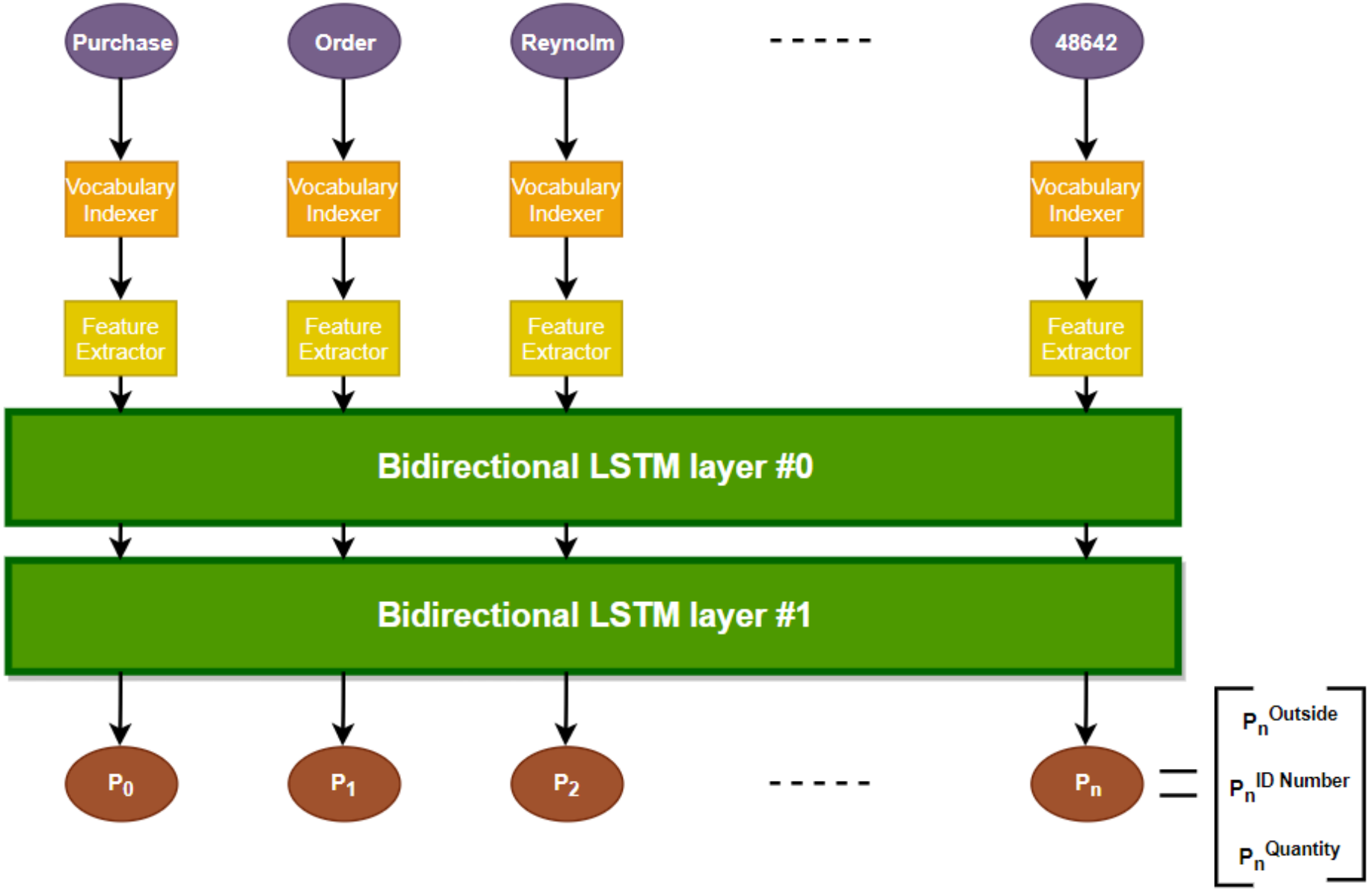
STEP #3: FEATURE EXTRACTOR



STEP #3: FEATURE EXTRACTOR



STEP #4: TOKEN LEVEL CLASSIFIER



BASELINE: FEEDFORWARD NEURAL NETWORK WITH LOCAL CONTEXT

- BLSTM layers are replaced by dense feed forward layers with equivalent number of parameters

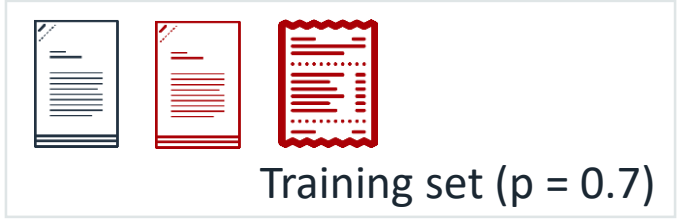
- For fair comparison, the baseline is enhanced with local context in its inputs:
  - For a given token, the feature vectors of its 4 closest tokens on the document - one in each direction - are concatenated with its own feature vector

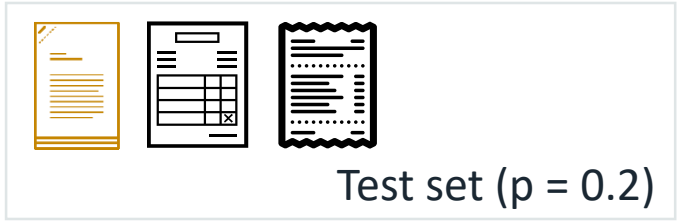
DATASET USED FOR EXPERIMENTS

- Dataset created with documents from Esker platform:
  - 28,570 purchase orders in English language
  - Originating from 2,818 distinct issuers, each having between 3 and 31 document instances
  - Within a document, there is on average more than 3 ordered products

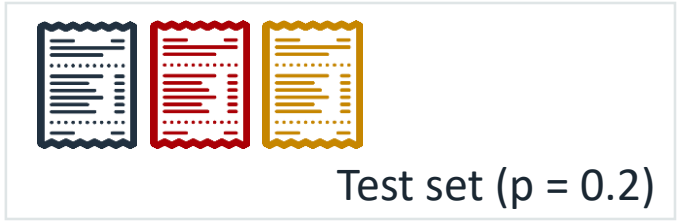
EXPERIMENTS DESCRIPTION

Splitting dataset at document level





Splitting dataset at issuer level



HYPER PARAMETERS VALUES BASED ON VALIDATION PERFORMANCES

- 64 dimensional textual embeddings
- RNN model is composed of two bidirectional LSTM layers of 1,300 cells each
- Baseline has two dense layers with 6,947 and 1,300 rectified linear units

EXPERIMENTS RESULTS

TABLE I
EXTRACTION PERFORMANCES FOR DOCUMENTLEVELSPLITTING EXPERIMENT.

	F1 score		Precision		Recall	
Field	Baseline	RNN	Baseline	RNN	Baseline	RNN
ID number	0.8532	**0.9064**	0.8633	**0.9083**	0.8435	**0.9051**
Quantity	0.9260	**0.9641**	0.9022	**0.9549**	0.9521	**0.9736**
Micro avg.	0.8889	**0.9343**	0.8825	**0.9309**	0.8956	**0.9380**

TABLE II
EXTRACTION PERFORMANCES FOR ISSUERLEVELSPLITTING EXPERIMENT.

	F1 score		Precision		Recall	
Field	Baseline	RNN	Baseline	RNN	Baseline	RNN
ID number	0.6851	**0.7522**	0.6887	**0.7694**	0.6870	**0.7380**
Quantity	0.8475	**0.8938**	0.8423	**0.9016**	0.8586	**0.8882**
Micro avg.	0.7640	**0.8207**	0.7625	**0.8338**	0.7692	**0.8102**

CONCLUSION

- Proposed a generic template-free approach for extracting table field in business documents that is able to deal with unknown layouts

- By experimenting on a large dataset of real world purchase orders, we showed that using recurrent connections is beneficial for the information extraction
  - On unknown templates, micro F1 score of 0.821 compared to 0.764 for the baseline

FUTURE WORK

- Experimenting model with character level textual representations generated by a RNN in order to provide a more granular parsing of the text

- Getting more structured predictions :
  - Conditional Random Field (CRF) layer on top of the RNN
  - Tackles end-to-end recognition of structured tabular entities

- Assess the behavior of our extraction models when confronted with a multilingual dataset

BIBLIOGRAPHY

1. D. Schuster, K. Muthmann, D. Esser, A. Schill, M. Berger, C. Weidling, K. Aliyev, and A. Hofmeier, “Intellix–end-user trained information extraction for document archiving,” in Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE, 2013, pp. 101–105.
2. M. Rusinol, T. Benkhelfallah, and V. Poulain dAndecy, “Field extraction from administrative documents by incremental structural templates,” in Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE, 2013, pp. 1100–1104.
3. D. Esser, D. Schuster, K. Muthmann, and A. Schill, “Few-exemplar information extraction for business documents.” in ICEIS (1), 2014, pp. 293–298.
4. D. Schuster, D. Esser, K. Muthmann, and A. Schill, “Modelspace cooperative document information extraction in flexible hierarchies.” in ICEIS (1), 2015, pp. 321–329.
5. V. P. d’Andecy, E. Hartmann, and M. Rusinol, “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.
6. Clinchant, S., Déjean, H., Meunier, J. L., Lang, E. M., & Kleber, F. (2018, April). Comparing Machine Learning Approaches for Table Recognition in Historical Register Books. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)* (pp. 133-138). IEEE.

BIBLIOGRAPHY

7. Schreiber, S., Agne, S., Wolf, I., Dengel, A., & Ahmed, S. (2017, November). Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on* (Vol. 1, pp. 1162-1167). IEEE.
8. Kavasidis, I., Palazzo, S., Spampinato, C., Pino, C., Giordano, D., Giuffrida, D., & Messina, P. (2018). A Saliency-based Convolutional Neural Network for Table and Chart Detection in Digitized Documents. *arXiv preprint arXiv:1804.06236*.
9. Zhu, G., Bethea, T. J., & Krishna, V. (2007, August). Extracting relevant named entities for automated expense reimbursement. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1004-1012). ACM.
10. Deckert, F., Seidler, B., Ebbecke, M., & Gillmann, M. (2011, September). Table content understanding in smartfix. In *2011 International Conference on Document Analysis and Recognition* (pp. 488-492). IEEE.
11. Y. Belaïd and A. Belaïd, “Morphological tagging approach in document analysis of invoices,” in Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 1. IEEE, 2004,pp. 469–472
12. Palm, R. B., Winther, O., & Laws, F. (2017, November). CloudScan-A configuration-free invoice analysis system using recurrent neural networks. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 406-413). IEEE.

BIBLIOGRAPHY

13. Chiu, J. P., & Nichols, E. (2015). Named entity recognition with bidirectional LSTM-CNNs. arXiv preprint arXiv:1511.08308.
14. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.
15. Yadav, V., & Bethard, S. (2018). A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2145-2158).
16. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.
17. R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in International Conference on Machine Learning, 2013, pp. 1310–1318.