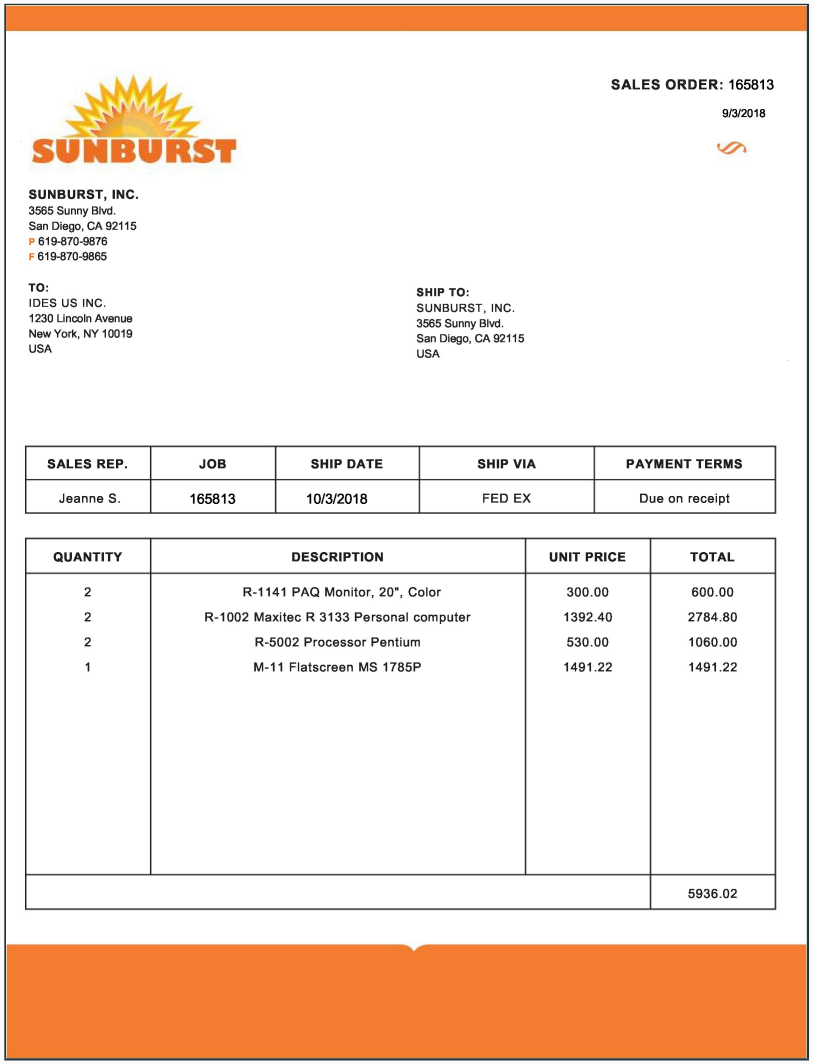
End-to-End Extraction of Structured Information from Business Documents with Pointer-Generator Networks

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

November 20, 2020

EMNLP 2020 Workshop on Structured Prediction for NLP

Information Extraction (IE) Task



SUNBURST

SALES ORDER: 165813

9/3/2018

SUNBURST, INC.
3565 Sunny Blvd.
San Diego, CA 92115
P 619-870-9876
F 619-870-9865

TO:
IDES US INC.
1230 Lincoln Avenue
New York, NY 10019
USA

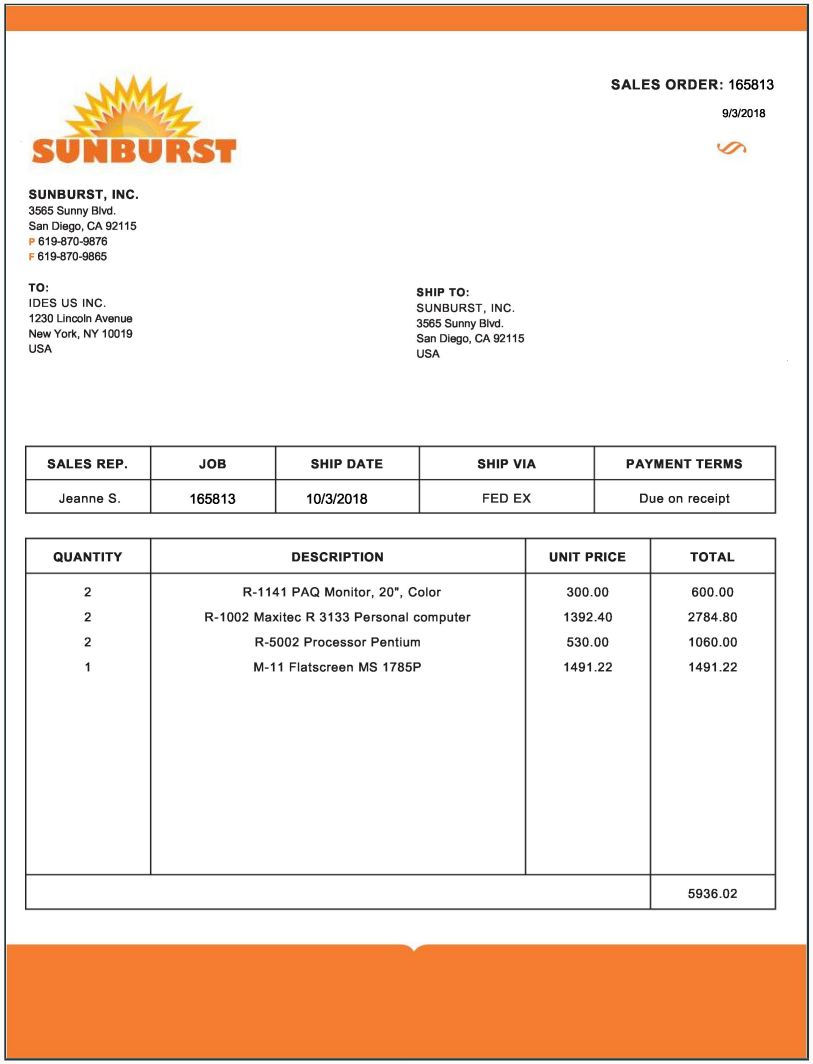
SHIP TO:
SUNBURST, INC.
3565 Sunny Blvd.
San Diego, CA 92115
USA

SALES REP.	JOB	SHIP DATE	SHIP VIA	PAYMENT TERMS
Jeanne S.	165813	10/3/2018	FED EX	Due on receipt

QUANTITY	DESCRIPTION	UNIT PRICE	TOTAL
2	R-1141 PAQ Monitor, 20", Color	300.00	600.00
2	R-1002 Maxitec R 3133 Personal computer	1392.40	2784.80
2	R-5002 Processor Pentium	530.00	1060.00
1	M-11 Flatscreen MS 1785P	1491.22	1491.22
			5936.02

Figure 1: A document sample and its extracted information.

Information Extraction (IE) Task



SUNBURST

SALES ORDER: 165813

9/3/2018

SUNBURST, INC.
3565 Sunny Blvd.
San Diego, CA 92115
P 619-870-9876
F 619-870-9865

TO:
IDES US INC.
1230 Lincoln Avenue
New York, NY 10019
USA

SHIP TO:
SUNBURST, INC.
3565 Sunny Blvd.
San Diego, CA 92115
USA

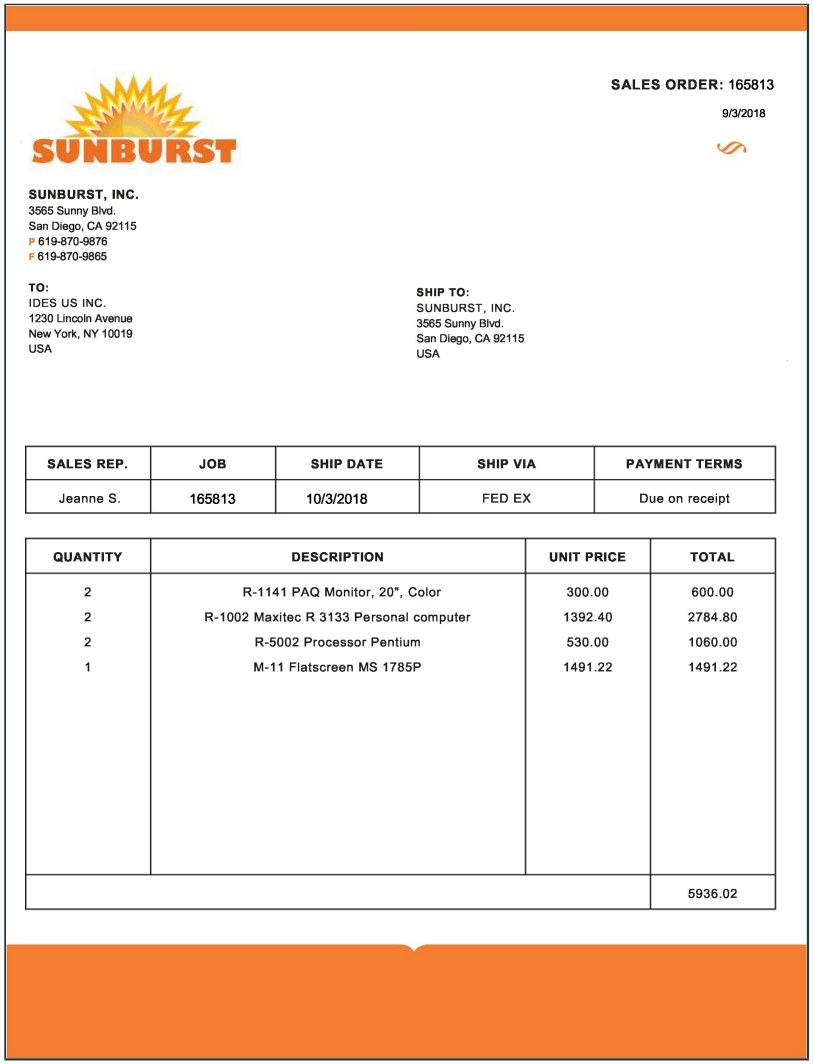
SALES REP.	JOB	SHIP DATE	SHIP VIA	PAYMENT TERMS
Jeanne S.	165813	10/3/2018	FED EX	Due on receipt

QUANTITY	DESCRIPTION	UNIT PRICE	TOTAL
2	R-1141 PAQ Monitor, 20", Color	300.00	600.00
2	R-1002 Maxitec R 3133 Personal computer	1392.40	2784.80
2	R-5002 Processor Pentium	530.00	1060.00
1	M-11 Flatscreen MS 1785P	1491.22	1491.22
			5936.02

```
<Extraction>
    <DocumentNumber>165813</DocumentNumber>
    <DocumentDate>2018-09-03</DocumentDate>
    <ShipDate>2018-10-03</ShipDate>
    <Total>5936.02</Total>
    <Product>
        <IDNumber>R-1141</IDNumber>
        <Quantity>2</Quantity>
    </Product>
    <Product>
        <IDNumber>R-1002</IDNumber>
        <Quantity>2</Quantity>
    </Product>
    <Product>
        <IDNumber>R-5002</IDNumber>
        <Quantity>2</Quantity>
    </Product>
    <Product>
        <IDNumber>M-11</IDNumber>
        <Quantity>1</Quantity>
    </Product>
</Extraction>
```

Figure 1: A document sample and its extracted information.

Information Extraction (IE) Task



SUNBURST

SALES ORDER: 165813

9/3/2018

SUNBURST, INC.
3565 Sunny Blvd.
San Diego, CA 92115
P 619-870-9876
F 619-870-9865

TO:
IDES US INC.
1230 Lincoln Avenue
New York, NY 10019
USA

SHIP TO:
SUNBURST, INC.
3565 Sunny Blvd.
San Diego, CA 92115
USA

SALES REP.	JOB	SHIP DATE	SHIP VIA	PAYMENT TERMS
Jeanne S.	165813	10/3/2018	FED EX	Due on receipt

QUANTITY	DESCRIPTION	UNIT PRICE	TOTAL
2	R-1141 PAQ Monitor, 20", Color	300.00	600.00
2	R-1002 Maxitec R 3133 Personal computer	1392.40	2784.80
2	R-5002 Processor Pentium	530.00	1060.00
1	M-11 Flatscreen MS 1785P	1491.22	1491.22
			5936.02

Unstructured fields

```
<Extraction>
    <DocumentNumber>165813</DocumentNumber>
    <DocumentDate>2018-09-03</DocumentDate>
    <ShipDate>2018-10-03</ShipDate>
    <Total>5936.02</Total>
    <Product>
        <IDNumber>R-1141</IDNumber>
        <Quantity>2</Quantity>
    </Product>
    <Product>
        <IDNumber>R-1002</IDNumber>
        <Quantity>2</Quantity>
    </Product>
    <Product>
        <IDNumber>R-5002</IDNumber>
        <Quantity>2</Quantity>
    </Product>
    <Product>
        <IDNumber>M-11</IDNumber>
        <Quantity>1</Quantity>
    </Product>
</Extraction>
```

Figure 2: Information may be unstructured: fields can be extracted independently.

Information Extraction (IE) Task

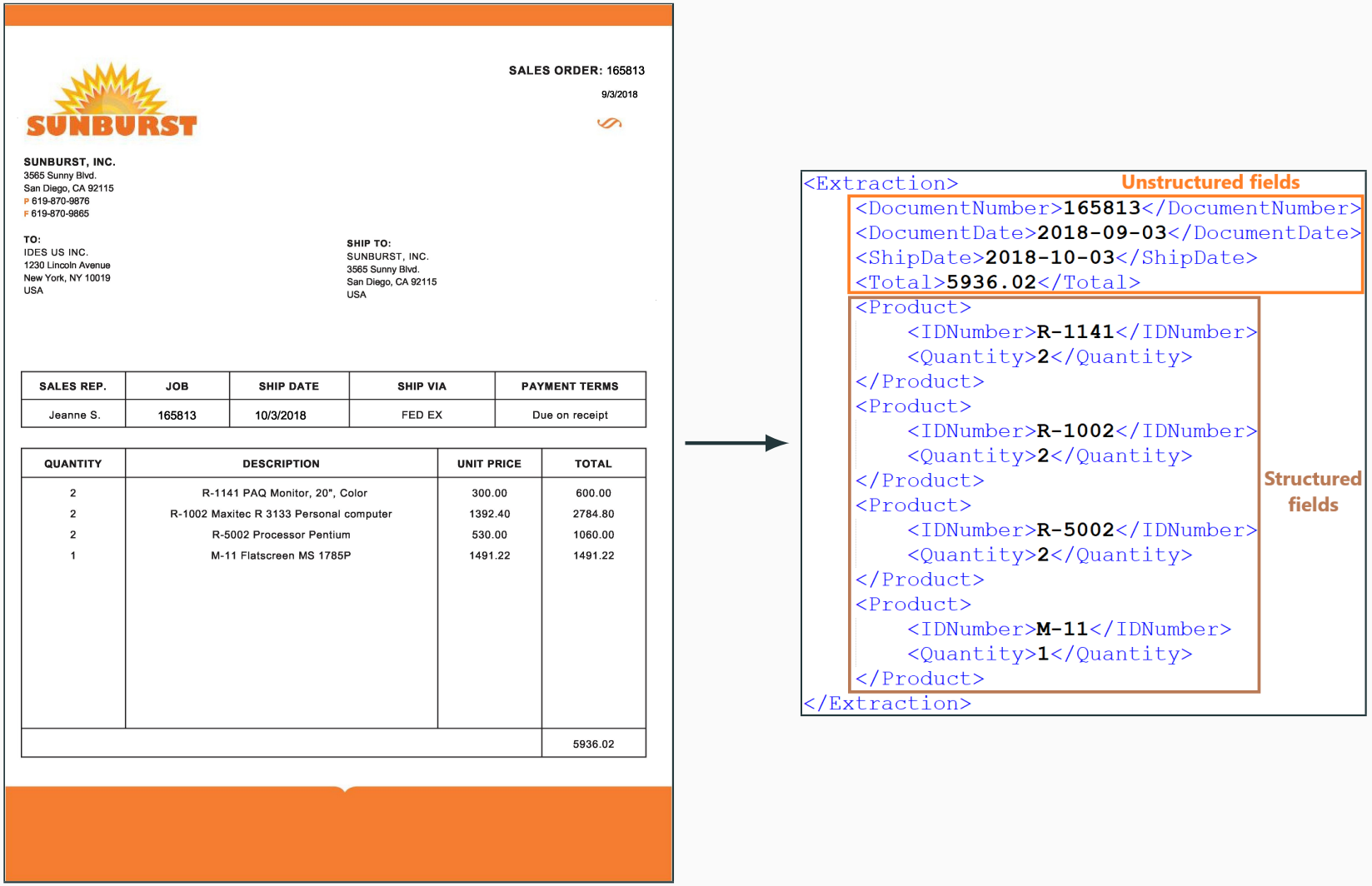


Figure 3: Information may be structured, e.g. tabular data.

Prior work in IE

- SOTA IE methods use neural models to deal with the variability of document layouts [1, 3, 4].

Prior work in IE

- SOTA IE methods use neural models to deal with the variability of document layouts [1, 3, 4].
- Vast majority of them are based on token classifiers that predict the information type of each token of the document.

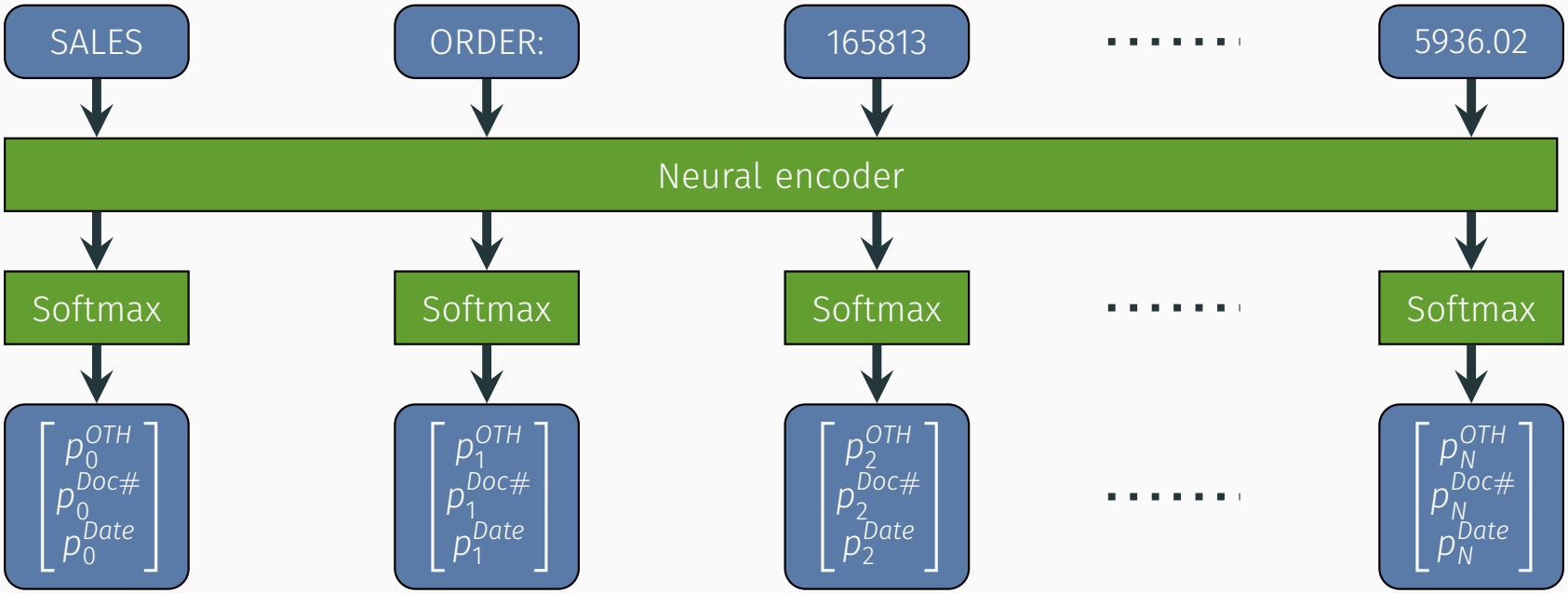


Figure 4: Token classifier for extracting the document number and date in Figure 1. The Other (*OTH*) class is dedicated to tokens which do not carry relevant information.

Drawbacks of token classifiers for IE

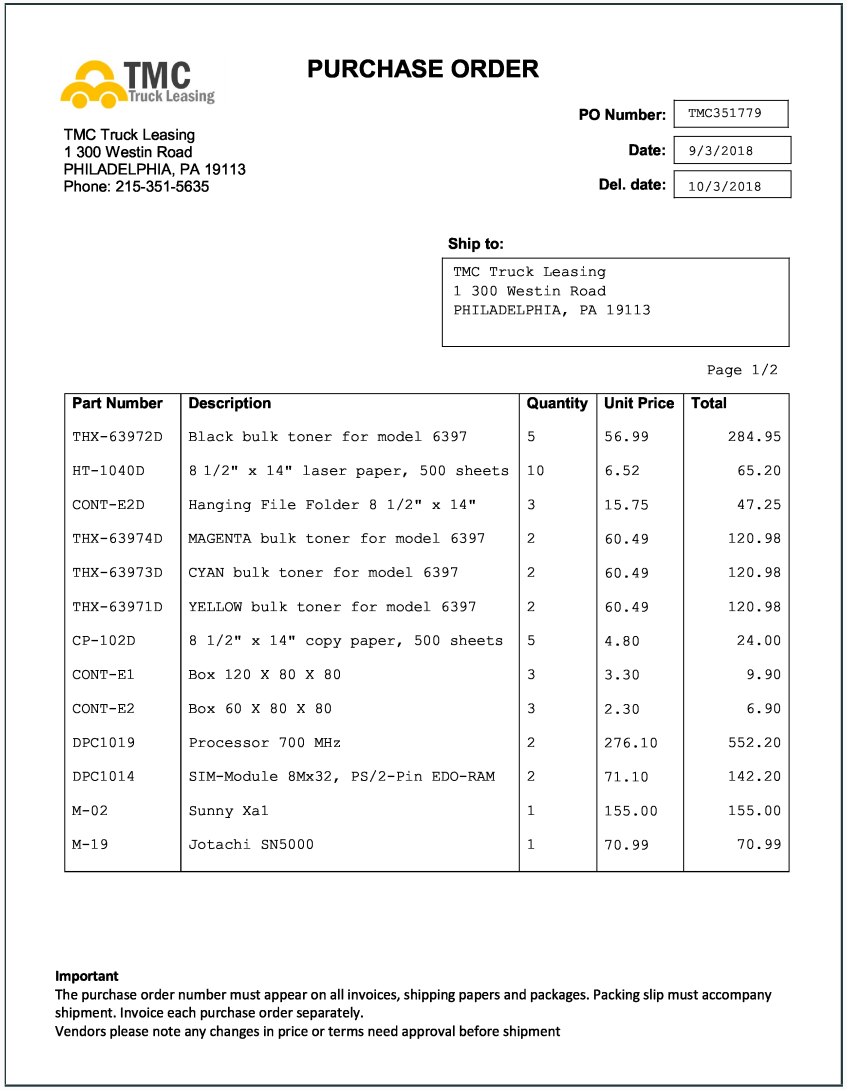
- Such approaches require token level supervision during training.

Drawbacks of token classifiers for IE

- Such approaches require token level supervision during training.
- Yet, deducing token labels from the extracted information is not trivial:

Drawbacks of token classifiers for IE

- Such approaches require token level supervision during training.
- Yet, deducing token labels from the extracted information is not trivial:
  - Multiple words of the document may share the same textual value while being semantically different.



TMC Truck Leasing

PURCHASE ORDER

TMC Truck Leasing
1 300 Westin Road
PHILADELPHIA, PA 19113
Phone: 215-351-5635

PO Number: TMC351779
Date: 9/3/2018
Del. date: 10/3/2018

Ship to:
TMC Truck Leasing
1 300 Westin Road
PHILADELPHIA, PA 19113

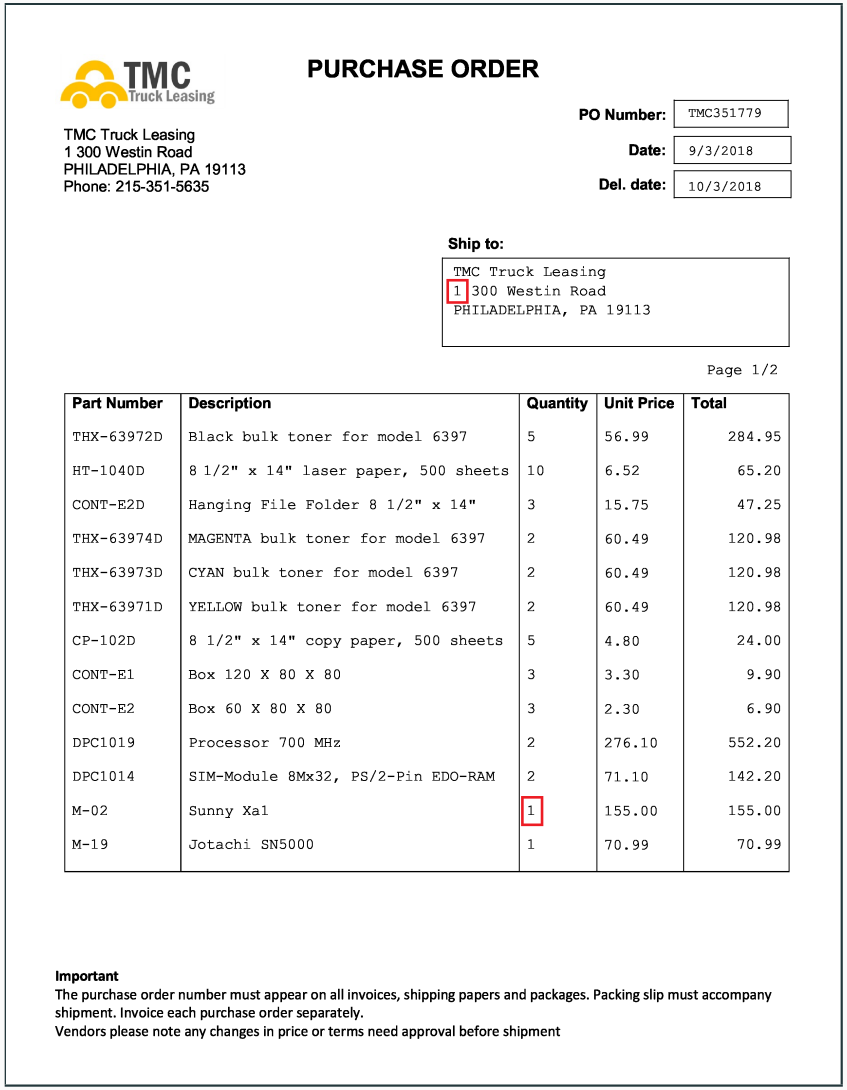
Page 1/2

Part Number	Description	Quantity	Unit Price	Total
THX-63972D	Black bulk toner for model 6397	5	56.99	284.95
HT-1040D	8 1/2" x 14" laser paper, 500 sheets	10	6.52	65.20
CONT-E2D	Hanging File Folder 8 1/2" x 14"	3	15.75	47.25
THX-63974D	MAGENTA bulk toner for model 6397	2	60.49	120.98
THX-63973D	CYAN bulk toner for model 6397	2	60.49	120.98
THX-63971D	YELLOW bulk toner for model 6397	2	60.49	120.98
CP-102D	8 1/2" x 14" copy paper, 500 sheets	5	4.80	24.00
CONT-E1	Box 120 X 80 X 80	3	3.30	9.90
CONT-E2	Box 60 X 80 X 80	3	2.30	6.90
DPC1019	Processor 700 MHz	2	276.10	552.20
DPC1014	SIM-Module 8Mx32, PS/2-Pin EDO-RAM	2	71.10	142.20
M-02	Sunny Xa1	1	155.00	155.00
M-19	Jotachi SN5000	1	70.99	70.99

Important
The purchase order number must appear on all invoices, shipping papers and packages. Packing slip must accompany shipment. Invoice each purchase order separately.
Vendors please note any changes in price or terms need approval before shipment

Drawbacks of token classifiers for IE

- Such approaches require token level supervision during training.
- Yet, deducing token labels from the extracted information is not trivial:
  - Multiple words of the document may share the same textual value while being semantically different.



TMC Truck Leasing

PURCHASE ORDER

TMC Truck Leasing
1 300 Westin Road
PHILADELPHIA, PA 19113
Phone: 215-351-5635

PO Number: TMC351779
Date: 9/3/2018
Del. date: 10/3/2018

Ship to:
TMC Truck Leasing
1 300 Westin Road
PHILADELPHIA, PA 19113

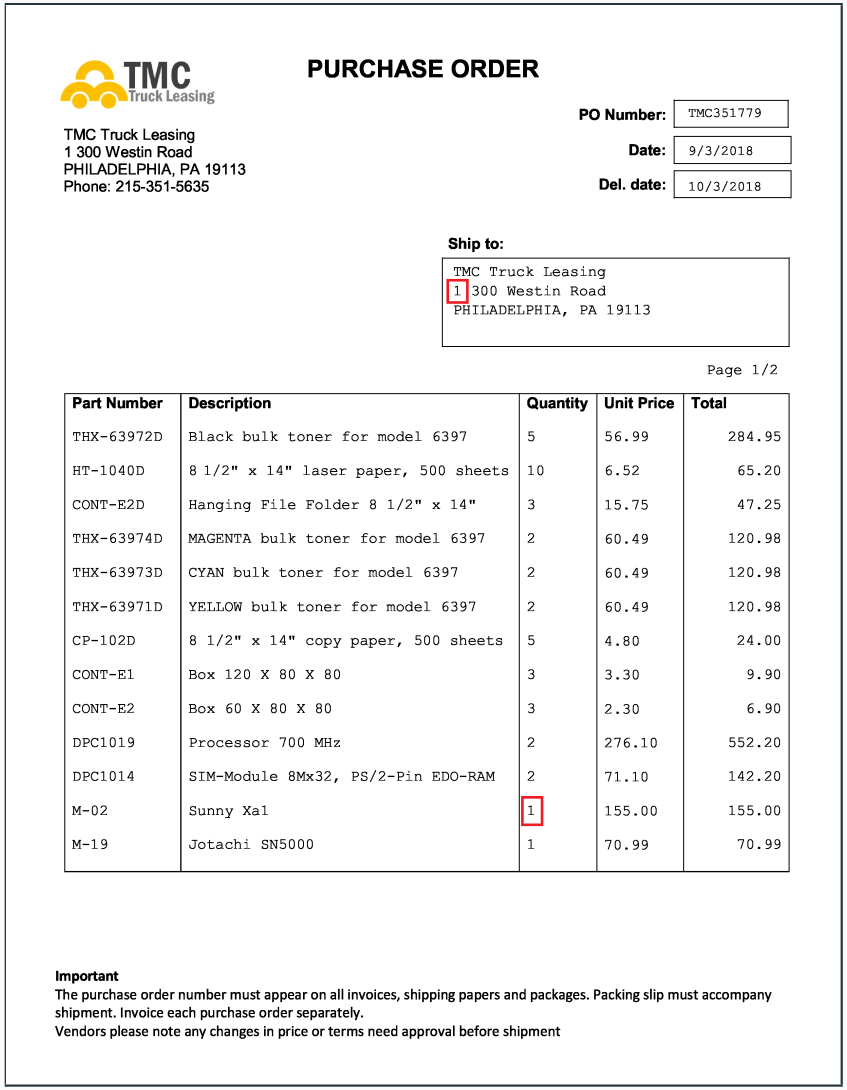
Page 1/2

Part Number	Description	Quantity	Unit Price	Total
THX-63972D	Black bulk toner for model 6397	5	56.99	284.95
HT-1040D	8 1/2" x 14" laser paper, 500 sheets	10	6.52	65.20
CONT-E2D	Hanging File Folder 8 1/2" x 14"	3	15.75	47.25
THX-63974D	MAGENTA bulk toner for model 6397	2	60.49	120.98
THX-63973D	CYAN bulk toner for model 6397	2	60.49	120.98
THX-63971D	YELLOW bulk toner for model 6397	2	60.49	120.98
CP-102D	8 1/2" x 14" copy paper, 500 sheets	5	4.80	24.00
CONT-E1	Box 120 X 80 X 80	3	3.30	9.90
CONT-E2	Box 60 X 80 X 80	3	2.30	6.90
DPC1019	Processor 700 MHz	2	276.10	552.20
DPC1014	SIM-Module 8Mx32, PS/2-Pin EDO-RAM	2	71.10	142.20
M-02	Sunny Xa1	1	155.00	155.00
M-19	Jotachi SN5000	1	70.99	70.99

Important
The purchase order number must appear on all invoices, shipping papers and packages. Packing slip must accompany shipment. Invoice each purchase order separately.
Vendors please note any changes in price or terms need approval before shipment

Drawbacks of token classifiers for IE

- Such approaches require token level supervision during training.
- Yet, deducing token labels from the extracted information is not trivial:
  - Multiple words of the document may share the same textual value while being semantically different.
- This often leads to overhead annotation costs.



TMC Truck Leasing

PURCHASE ORDER

TMC Truck Leasing
1 300 Westin Road
PHILADELPHIA, PA 19113
Phone: 215-351-5635

PO Number: TMC351779
Date: 9/3/2018
Del. date: 10/3/2018

Ship to:
TMC Truck Leasing
1 300 Westin Road
PHILADELPHIA, PA 19113

Page 1/2

Part Number	Description	Quantity	Unit Price	Total
THX-63972D	Black bulk toner for model 6397	5	56.99	284.95
HT-1040D	8 1/2" x 14" laser paper, 500 sheets	10	6.52	65.20
CONT-E2D	Hanging File Folder 8 1/2" x 14"	3	15.75	47.25
THX-63974D	MAGENTA bulk toner for model 6397	2	60.49	120.98
THX-63973D	CYAN bulk toner for model 6397	2	60.49	120.98
THX-63971D	YELLOW bulk toner for model 6397	2	60.49	120.98
CP-102D	8 1/2" x 14" copy paper, 500 sheets	5	4.80	24.00
CONT-E1	Box 120 X 80 X 80	3	3.30	9.90
CONT-E2	Box 60 X 80 X 80	3	2.30	6.90
DPC1019	Processor 700 MHz	2	276.10	552.20
DPC1014	SIM-Module 8Mx32, PS/2-Pin EDO-RAM	2	71.10	142.20
M-02	Sunny Xa1	1	155.00	155.00
M-19	Jotachi SN5000	1	70.99	70.99

Important
The purchase order number must appear on all invoices, shipping papers and packages. Packing slip must accompany shipment. Invoice each purchase order separately.
Vendors please note any changes in price or terms need approval before shipment

End-to-End (E2E) Information Extraction

- Palm et al. [2] propose a model that directly output the textual values of fields but the approach is designed only for unstructured fields.

End-to-End (E2E) Information Extraction

- Palm et al. [2] propose a model that directly output the textual values of fields but the approach is designed only for unstructured fields.
- In this work, we introduce an E2E method that is also suitable for structured information.

Pointer-Generator Networks (PGN) [5] for IE

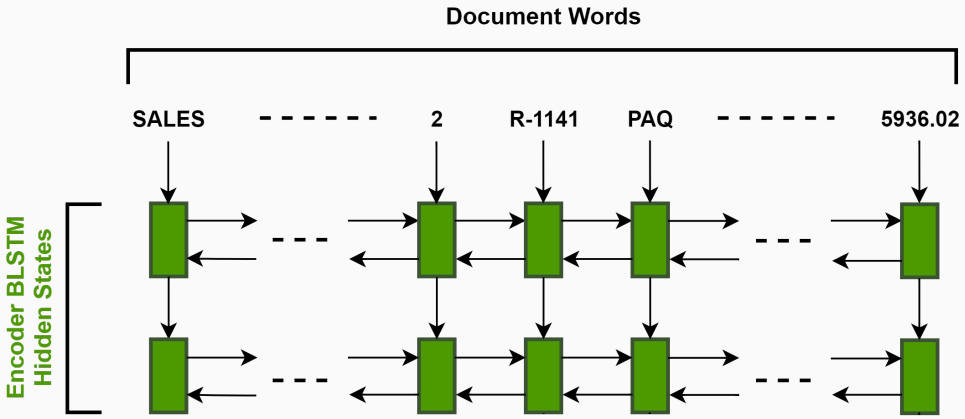


Figure 5: Our PGN extracting information of the document in Figure 1

Pointer-Generator Networks (PGN) [5] for IE

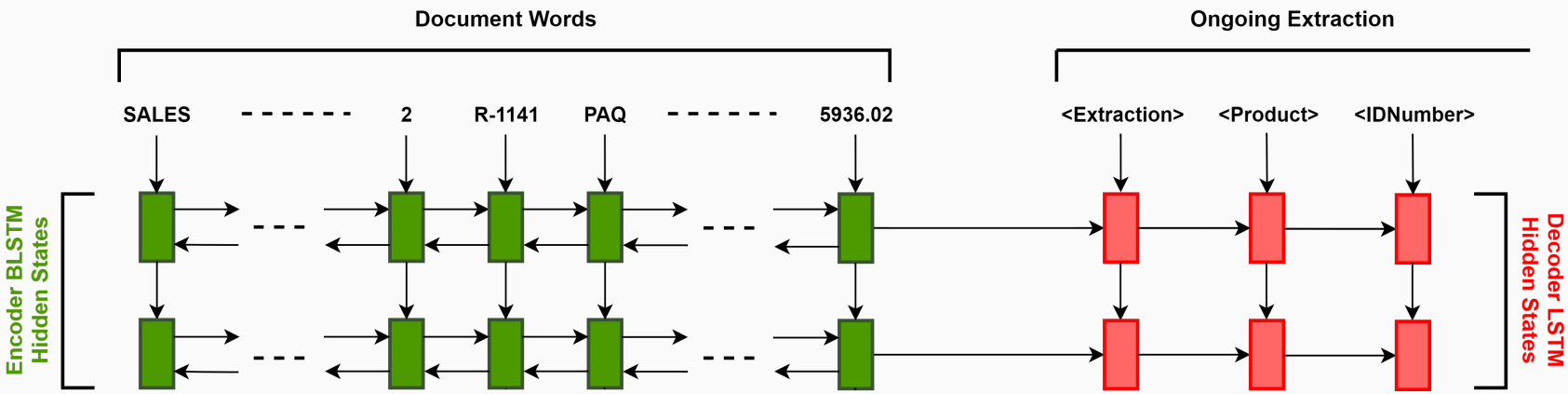


Figure 5: Our PGN extracting information of the document in Figure 1

Pointer-Generator Networks (PGN) [5] for IE

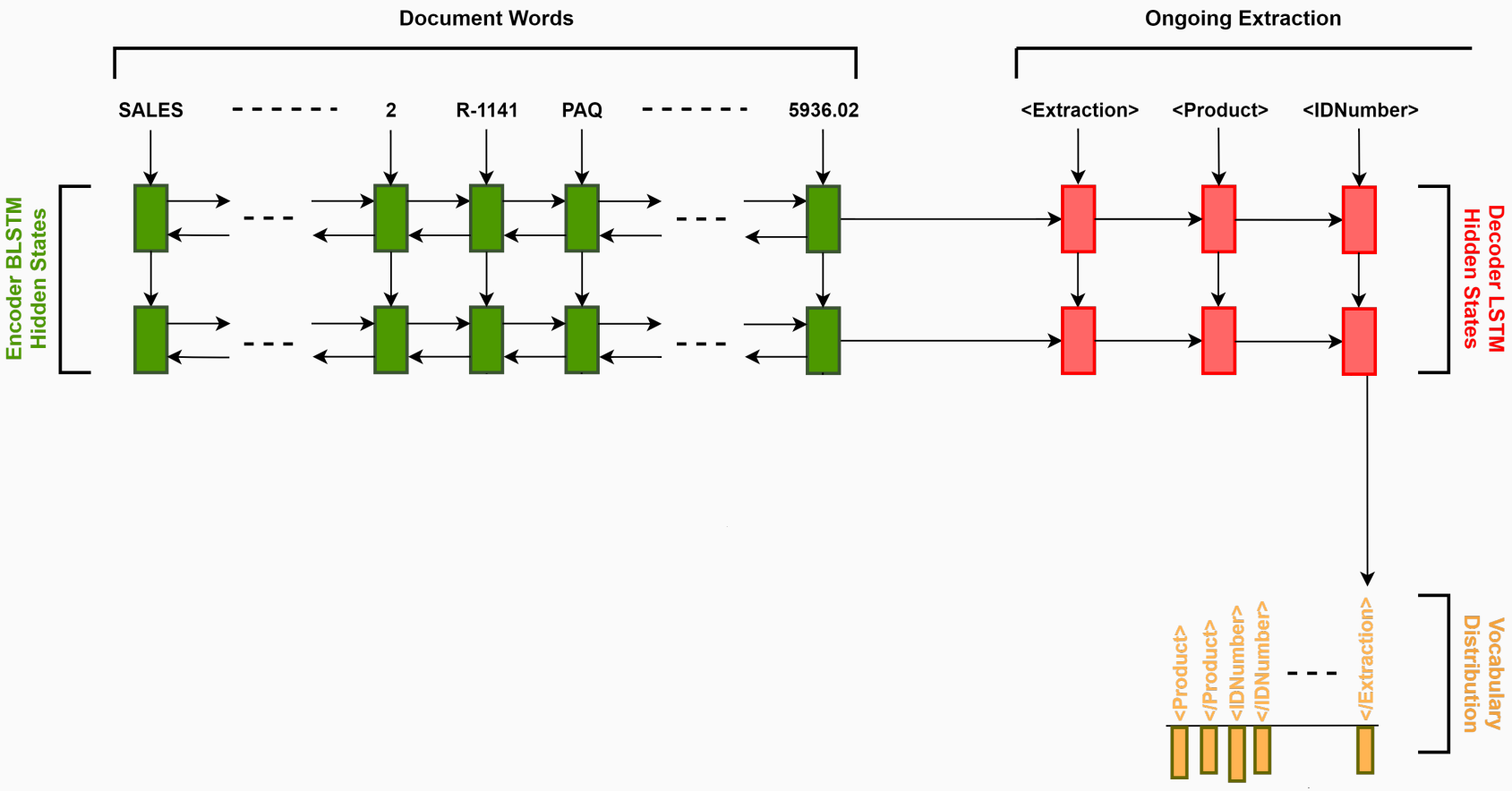


Figure 5: Our PGN extracting information of the document in Figure 1

Pointer-Generator Networks (PGN) [5] for IE

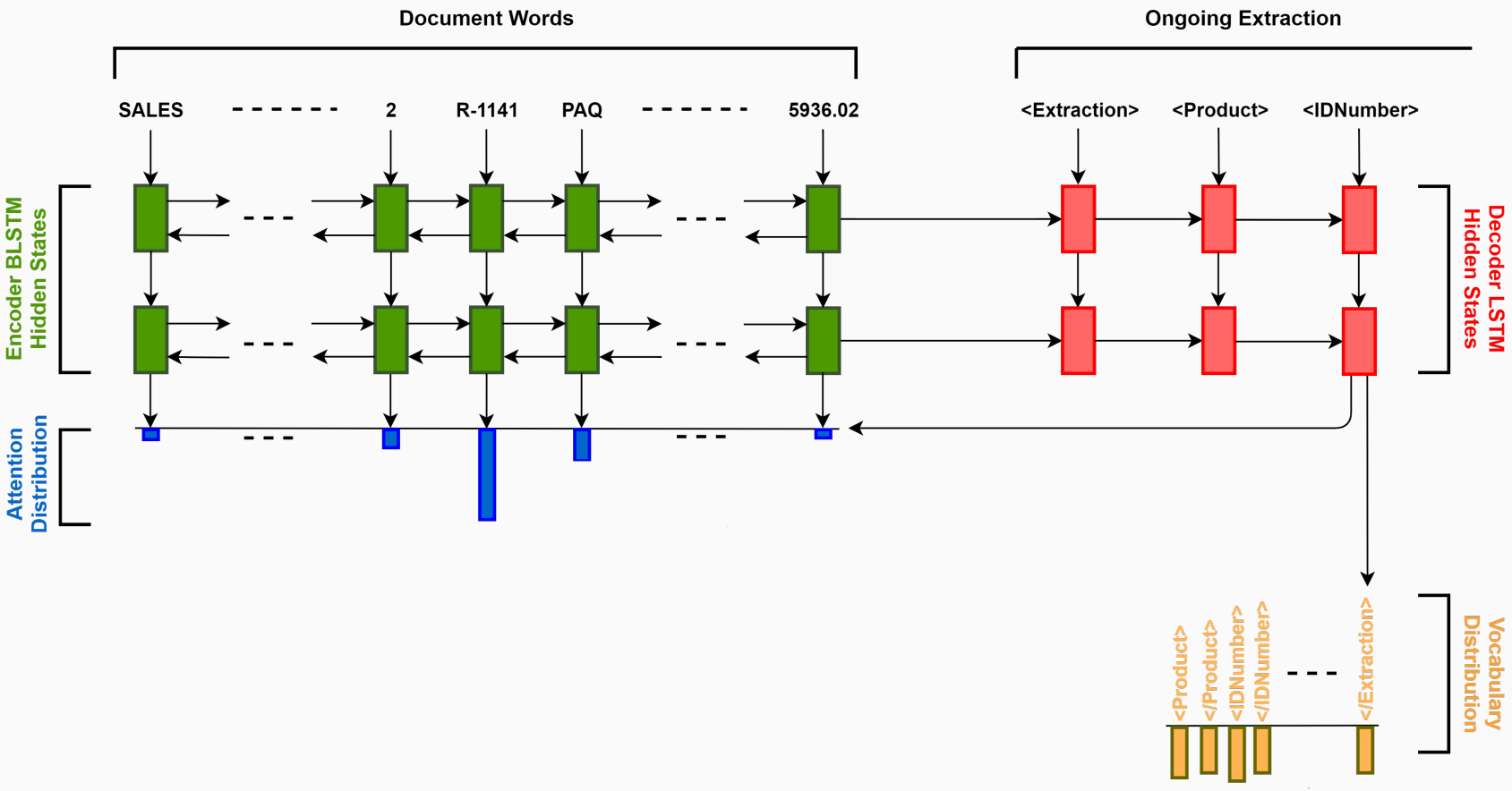


Figure 5: Our PGN extracting information of the document in Figure 1

Pointer-Generator Networks (PGN) [5] for IE

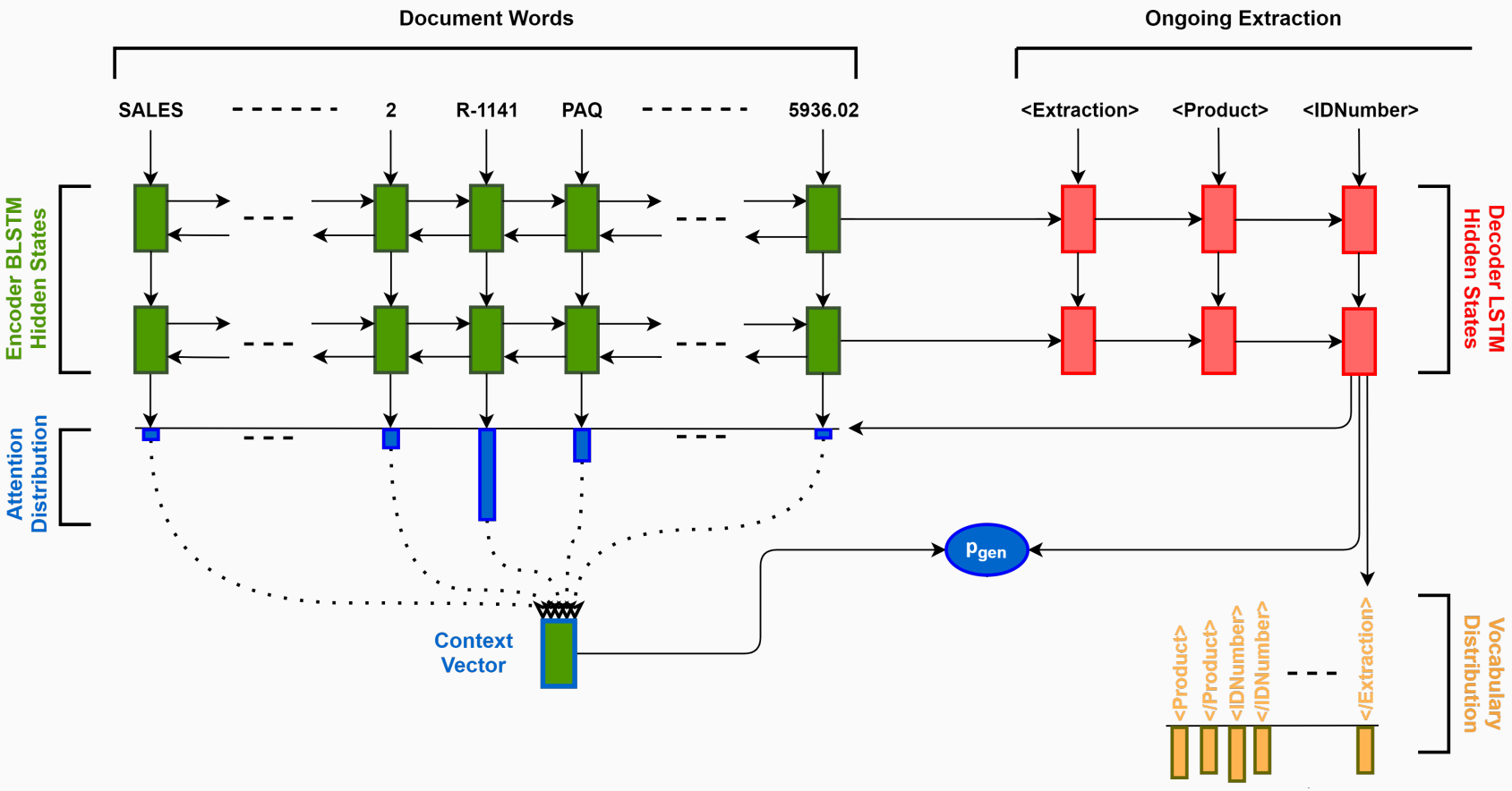


Figure 5: Our PGN extracting information of the document in Figure 1

Pointer-Generator Networks (PGN) [5] for IE

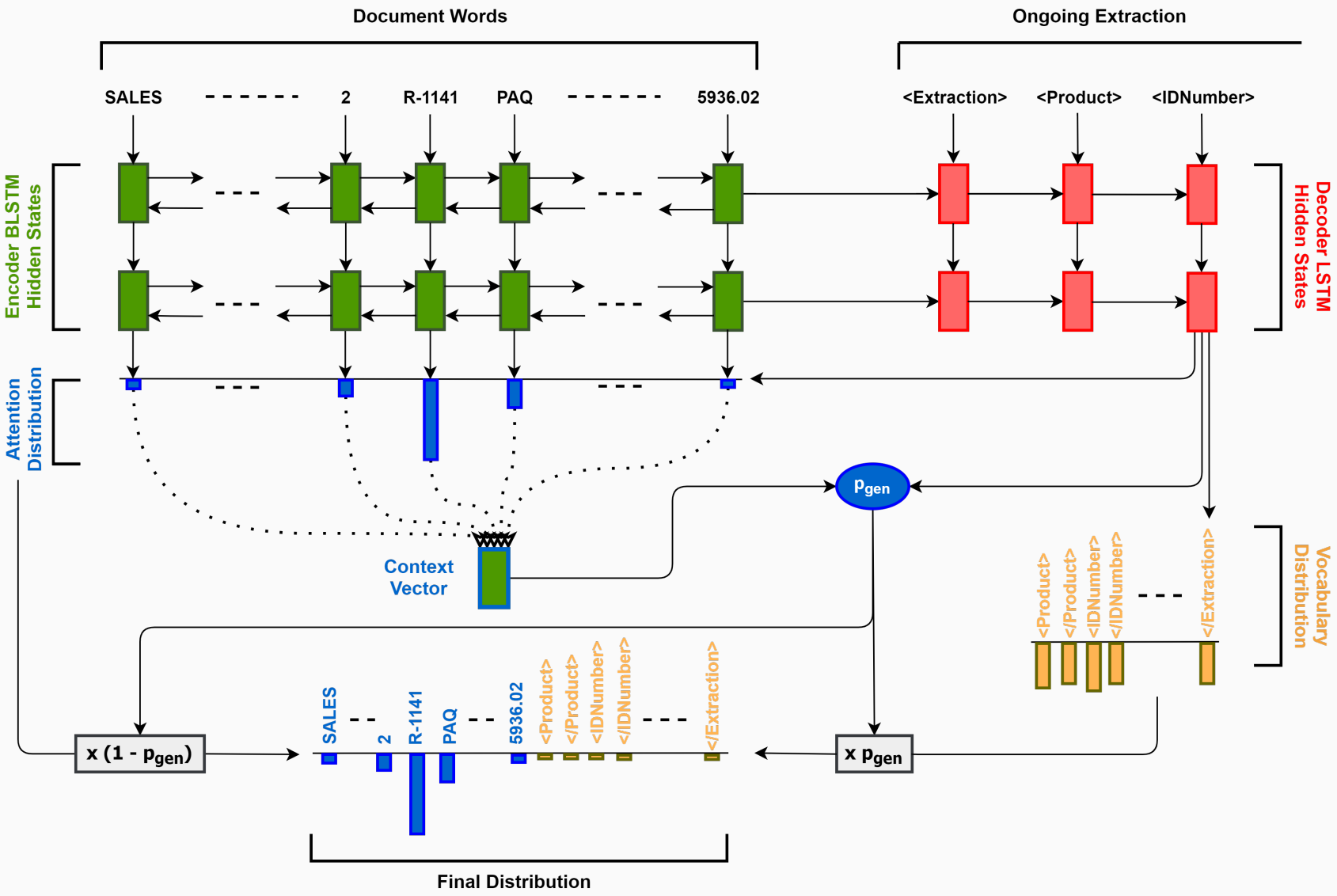


Figure 5: Our PGN extracting information of the document in Figure 1

Dataset used for evaluation

- Private real-world dataset of 219k purchase orders.

Dataset used for evaluation

- Private real-world dataset of 219k purchase orders.
- Extraction of their ordered products with two fields:

Dataset used for evaluation

- Private real-world dataset of 219k purchase orders.
- Extraction of their ordered products with two fields:

```
<Extraction>
    <Product>
        <IDNumber>R-1141</IDNumber>
        <Quantity>2</Quantity>
    </Product>
    <Product>
        <IDNumber>R-1002</IDNumber>
        <Quantity>2</Quantity>
    </Product>
    <Product>
        <IDNumber>R-5002</IDNumber>
        <Quantity>2</Quantity>
    </Product>
    <Product>
        <IDNumber>M-11</IDNumber>
        <Quantity>1</Quantity>
    </Product>
</Extraction>
```

Post-processing gain metric

For a given field:

$$1 - \frac{\#\text{ deletions} + \#\text{ insertions} + \#\text{ modifications}}{N}$$

where N is the number of ground truth instances in the document for this field.

Results

Table 1: Post-processing gains when extracting the products from the test documents. *% Perfect* column indicates the percentage of documents perfectly processed by each model.

	ID number	Quantity	Micro avg.	% Perfect
Word classifier	**0.754**	**0.855**	**0.804**	67.4
PGN	0.711	0.832	0.771	**68.2**

Summary of contributions

We proposed a new end-to-end method for extracting structured information that:

Summary of contributions

We proposed a new end-to-end method for extracting structured information that:

- is competitive with traditional IE methods.

Summary of contributions

We proposed a new end-to-end method for extracting structured information that:

- is competitive with traditional IE methods.
- can significantly reduce annotation efforts.

Summary of contributions

We proposed a new end-to-end method for extracting structured information that:

- is competitive with traditional IE methods.
- can significantly reduce annotation efforts.
- is flexible to extract any information types.

Thank you !

Paper & contact information

Seeking for more details about our work ? Click here for reaching our paper.

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

clement.sage@liris.cnrs.fr

References i

[1] A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. Höhne, and J. B. Faddoul.
Chargrid: Towards understanding 2d documents.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4459–4469, 2018.

[2] R. B. Palm, F. Laws, and O. Winther.
Attend, copy, parse end-to-end information extraction from documents.
In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 329–336. IEEE, 2019.

References ii

[3] 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)*, pages 406–413. IEEE, 2017.

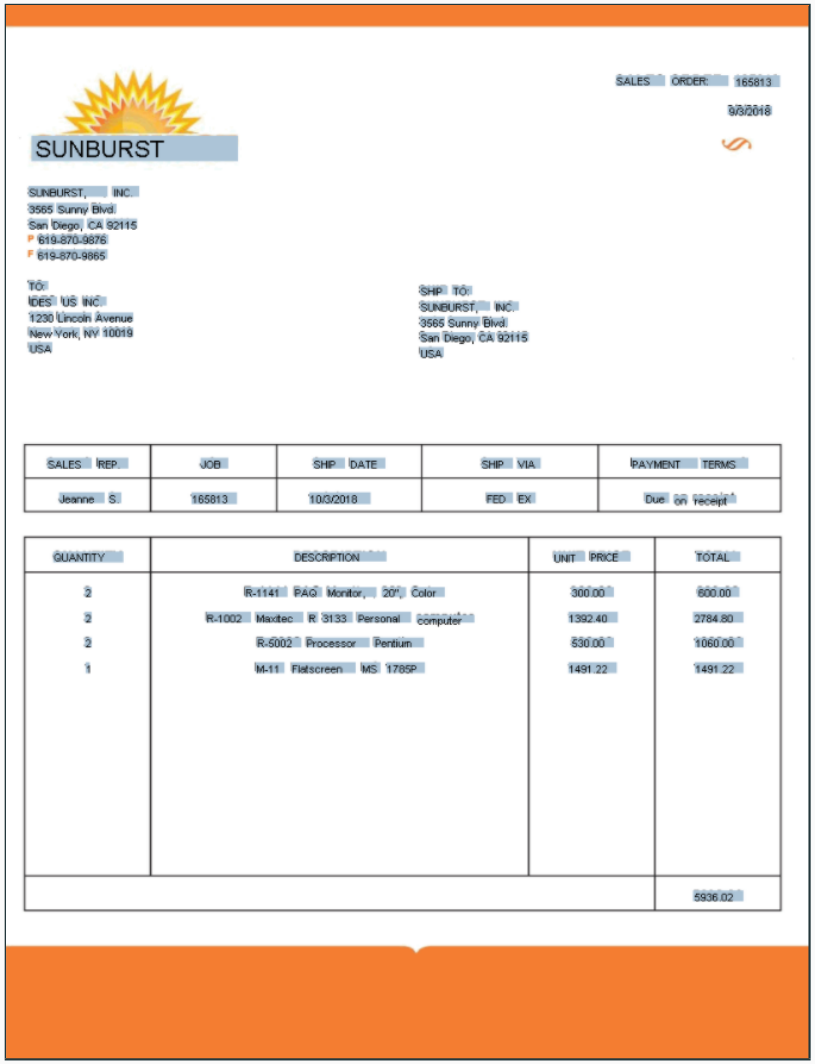
[4] C. Sage, A. Aussem, H. Elghazel, V. Eglin, and J. Espinas.
Recurrent neural network approach for table field extraction in business documents.
In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1308–1313, 2019.

References iii

[5] A. See, P. J. Liu, and C. D. Manning.
Get to the point: Summarization with pointer-generator networks.
In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, 2017.

Supplementary materials

Information Extraction (IE) Task




```
<Extraction>
    <DocumentNumber>165813</DocumentNumber>
    <DocumentDate>2018-09-03</DocumentDate>
    <ShipDate>2018-10-03</ShipDate>
    <Total>5936.02</Total>
    <Product>
        <IDNumber>R-1141</IDNumber>
        <Quantity>2</Quantity>
    </Product>
    <Product>
        <IDNumber>R-1002</IDNumber>
        <Quantity>2</Quantity>
    </Product>
    <Product>
        <IDNumber>R-5002</IDNumber>
        <Quantity>2</Quantity>
    </Product>
    <Product>
        <IDNumber>M-11</IDNumber>
        <Quantity>1</Quantity>
    </Product>
</Extraction>
```


Figure 6: We assume that the text of the document has been retrieved before extracting its information.

Word representations

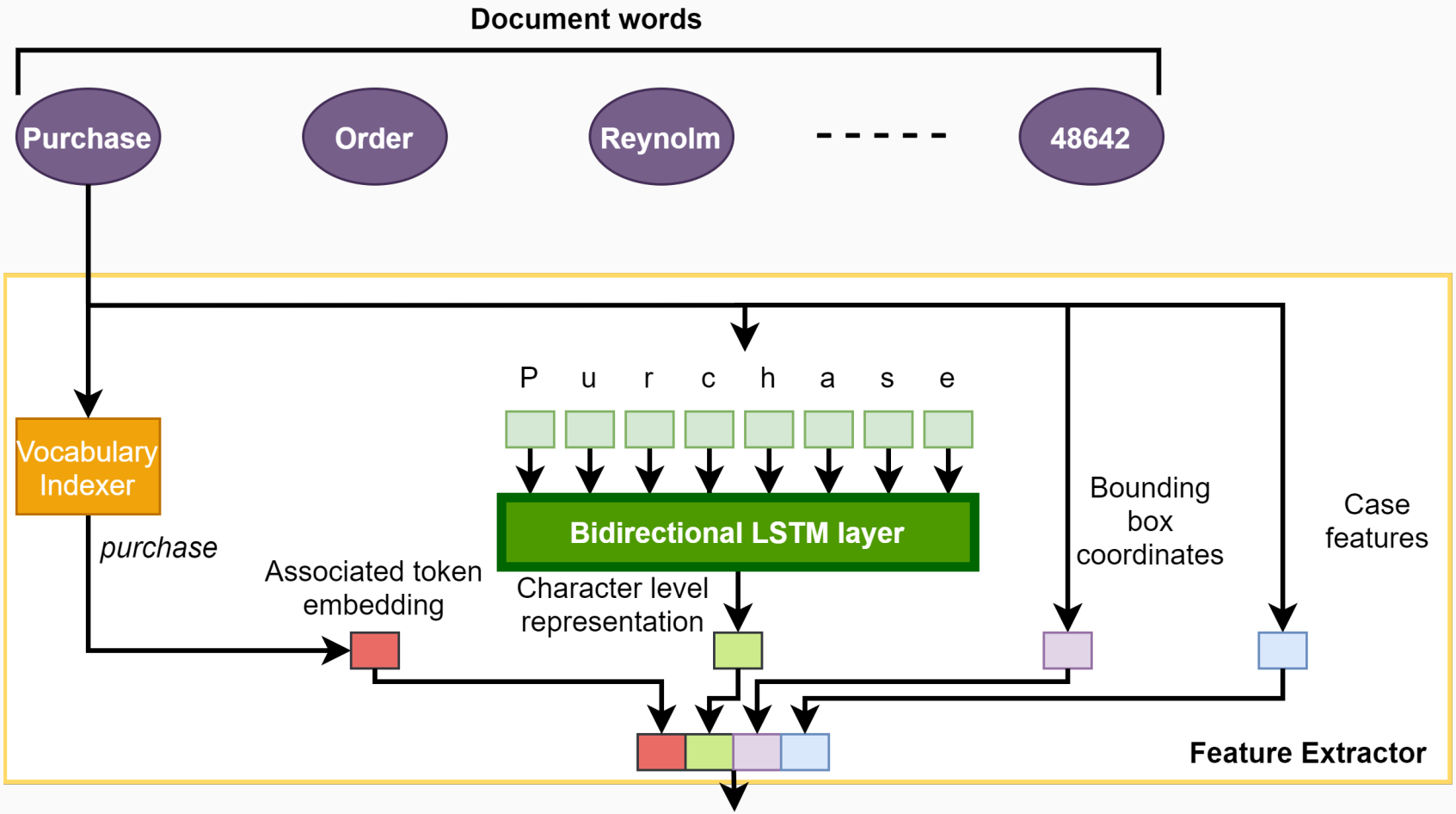


Figure 7: The word representations are made of fine-grained textual features and spatial coordinates.

Visualizing the attention mechanism

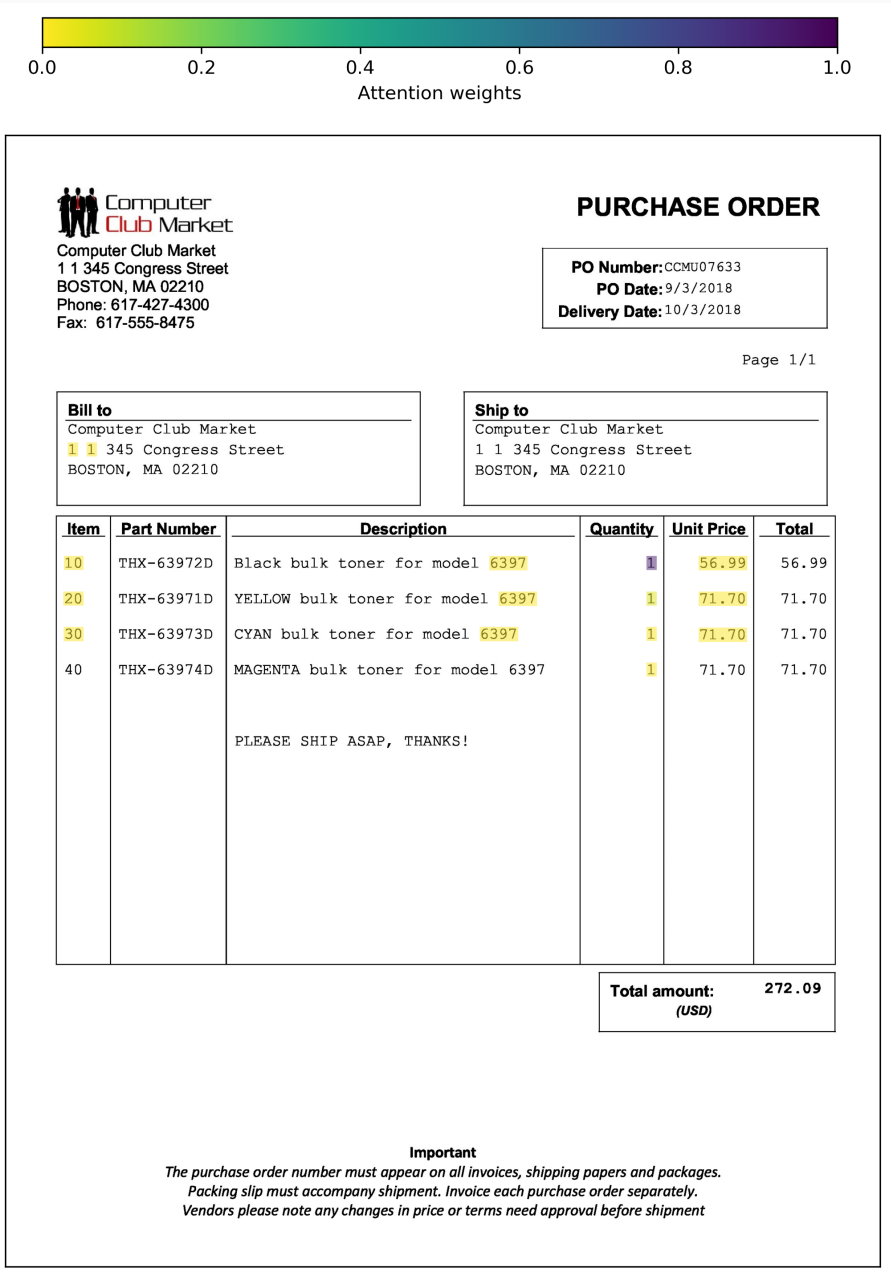


Figure 8: Attention weights for the top-15 document words at the 6^{th} time step. The PGN has previously outputted the tokens *<Product>*, *<IDNumber>*, *THX-63972D*, *</IDNumber>* and *<Quantity>*.

Further quantitative results

Table 2: Micro averaged gains over the test set conditioned on the number N of products in the document.

	$N \leq 3$	$3 < N < 15$	$N \geq 15$
Documents	33,332	7,820	1,613
Product entities	46,893	53,771	44,094
Word classifier	0.804	**0.807**	**0.801**
PGN	**0.820**	0.791	0.696

Presentation theme

Get the source of this theme and the demo presentation from

github.com/matze/mtheme