(047)

Information Extraction from Electricity Invoices through Named Entity Recognition with Transformers

A. Salgado and J. Sánchez
University of Las Palmas de Gran Canaria, Computer Science Department,
35017 Las Palmas de Gran Canaria, Spain
E-mail: {agustin.salgado, jsanchez}@ulpgc.es

Summary: This article describes a method for automatically extracting information from electricity invoices. This type of documents contains rich information about the billing of each supply point and data about the customer, the contract, or the electricity company. In this work, we train a neural network to classify the input data among eighty-six different labels. We use the IDSEM dataset that contains 75.000 electricity invoices of the Spanish electricity market in PDF format. Each document is converted into text format and the classification is carried out through a named entity recognition (NER) process. The underlying neural network used in the process is a Transformer. The results demonstrate that the proposed method correctly classifies the majority of the labels with high accuracy. Furthermore, the method exhibits robustness in handling invoices with different layouts and contents, highlighting its versatility and reliability.

Keywords: Machine learning, Natural Language Processing (NLP), Named entity recognition, Transformer, Electricity invoices.

1. Introduction

Electricity invoices are complex documents that contain a large variety of information about the electricity consumption, the billing, the customer, or the electricity company. Many companies need to incorporate the data into their information systems, but this is not an easy task. Information extraction from semi-structured documents plays a significant role in processing this type of invoices fast and reliably.

This task is important for both customers and utility companies. The goal is to automatically extract the information that provides better customer service and optimize operations.

The main impediment to develop new models for extracting information from these bills is the lack of high-quality datasets. In this respect, the recently proposed IDSEM [14] dataset was built to solve this problem, containing many electricity invoices in PDF format of the Spanish market. This is an interesting dataset that has many samples, with many different layouts, where the contents were randomly generated to avoid data protection issues. There are some other datasets about electricity, such as [1, 7 or 16], but these do not completely serve for our purposes since they only contain information about household electricity consumption.

One of the challenges in information extraction from electricity bills is the diversity of formats and layouts, given by the large number of companies that operate in the market. This makes it necessary to rely on deep learning techniques for the development of a general-purpose system.

In this work, we propose a new method for automatically extracting information from electricity invoices. We use the IDSEM database that contains 75.000 invoices with eighty-six different labels. It has a total of nine templates from different electricity companies, each one with its own layout and contents.

We develop a pipeline to process the information, where the bills are first converted into text, and then tokenized and passed through a named entity recognition (NER) system to detect the entities. The NER relies on a Transformer [18], which can detect most of the entities with high accuracy. Since the entities may appear multiple times in various parts of the documents, we fuse the information to obtain a unique value per label. The NER system is an important component of many NLP applications, and it is typically used for different tasks, such as information extraction, question answering or text classification.

Transformers are a powerful type of neural network architecture that has been very successful in NLP. They are particularly effective at capturing long-range dependencies and contextual relationships in text data. However, the text used in invoices is more structured and contains diverse data, that mixes different types of amounts (for example, bill prices and electricity consumption), with textual information. Therefore, there is little contextual information in some cases.

We review some of the most important techniques for this problem in Section 2. Section 3 details the features of the IDSEM dataset and the method that we have designed for recognizing the contents of the invoices. The experimental results are given in Section 4, with a focus on the outputs of the different steps of the pipeline and a study on the performance of the method. Finally, some concluding remarks in Section 5.

2. Related Work

The field of information extraction from semi-structured documents has witnessed significant advancements during the last years. Extracting relevant data from invoices or receipts poses a unique challenge due to the variability in formats and layouts.

Many effective methods have been developed to tackle this problem. Some of them are based on template matching [2, 12], which involves creating predefined templates that capture the structure and key fields of the documents. These templates serve as reference points for information extraction. However, these types of methods are highly dependent on accurately matching document layouts to predefined templates and may not handle variations well.

Rule-based extraction methods (SmartFix [3, 4, 8], Intellix [15]), on the other hand, rely on defining a set of rules or patterns to identify and extract relevant information from the documents. These rules can be based on regular expressions, keywords, or positional patterns. While rule-based methods are straight forward to implement, they can be brittle and require regular maintenance to accommodate document variations.

Traditional machine learning approaches have also been employed to extract information from invoices. Supervised learning involves training a model on annotated data to recognize patterns and extract relevant information. Unsupervised learning, on the other hand, leverages clustering algorithms to identify similar document structures and extract information accordingly.

Natural Language Processing (NLP) techniques treat documents as text and include several tasks, such as named entity recognition [5], part-of-speech tagging, and semantic parsing, which are utilized to extract structured information from unstructured text. These techniques enable the identification of entities, relationships, and key attributes within the document text.

Deep Learning models [9] have shown promising results for information extraction in recent years. Convolutional Neural Networks (CNNs) [6, 19], Recurrent Neural Networks (RNNs) [10, 13], or Transformer-based architectures [20] can be applied for tasks like document layout analysis, character recognition, and sequence labeling.

Hybrid approaches combine multiple methods, such as rule-based and machine learning techniques, to achieve more accurate and robust extraction. For example, rule-based methods can be used for initial extraction, followed by machine learning models to handle variations, and improve extraction accuracy.

The effectiveness of these methods depends on the availability of high-quality training data, domain-specific knowledge, and the complexity of the documents being processed. Additionally, ongoing model evaluation and adaptation are crucial to maintain accurate extraction performance as document formats evolve.

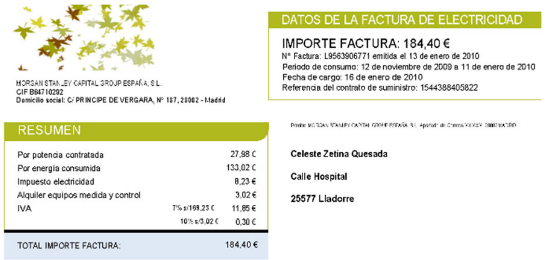
3. Material and Methods

3.1. Electricity Invoices Dataset

We use the IDSEM [14] dataset that contains electricity invoices in PDF format. This recently published database is organized in two directories with 30.000 files for training and 45.000 for testing. Each bill is defined by eighty-six different labels that are stored in JSON files. In the training set, there are two PDF files: one for the invoice and another with annotations around the labels to facilitate information extraction during the training process. The label codes are placed at the beginning and end of the corresponding values, such as #J5 191.32 #J5, where J5 represents the code for the "total price".

For each invoice, there is a corresponding JSON file that contains the value of each tag embedded in the PDF. This JSON file is used to verify the accuracy of the information extracted from the original bill.

Fig. 1 displays a fragment of an invoice from the IDSEM database and Fig. 2 shows the same fragment with label annotations.



MORGAN STANLEY CAPITAL GROUP ESPAÑA, S.L.
CIF B84710292
Domicilio social: C/ PRINCIPE DE VERGARA, Nº 187, 28002 - Madrid

DATOS DE LA FACTURA DE ELECTRICIDAD

IMPORTE FACTURA: 184,40 €
Nº Factura: L9563906771 emitida el 13 de enero de 2010
Periodo de consumo: 12 de noviembre de 2009 a 11 de enero de 2010
Fecha de cargo: 16 de enero de 2010
Referencia del contrato de suministro: 1544388405822

RESUMEN

Por potencia contratada		27,98 €
Por energía consumida		133,02 €
Impuesto electricidad		8,23 €
Alquiler equipos medida y control		3,02 €
IVA	7% s/169,23 €	11,85 €
	10% s/3,02 €	0,30 €
TOTAL IMPORTE FACTURA:		184,40 €

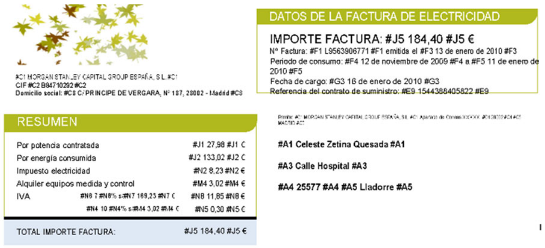
Remite: MORGAN STANLEY CAPITAL GROUP ESPAÑA, S.L. Apartado de Correos XXXXX 28002 MADRID

Celeste Zetina Quesada

Calle Hospital

25577 Lladorre

Fig. 1. A fragment of an invoice of the IDSEM dataset.



#C1 MORGAN STANLEY CAPITAL GROUP ESPAÑA, S.L. #C1
CIF #C2 B84710292 #C2
Domicilio social: #C8 C/ PRINCIPE DE VERGARA, Nº 187, 28002 - Madrid #C8

DATOS DE LA FACTURA DE ELECTRICIDAD

IMPORTE FACTURA: #J5 184,40 #J5 €
Nº Factura: #F1 L9563906771 #F1 emitida el #F3 13 de enero de 2010 #F3
Periodo de consumo: #F4 12 de noviembre de 2009 #F4 a #F5 11 de enero de 2010 #F5
Fecha de cargo: #G3 16 de enero de 2010 #G3
Referencia del contrato de suministro: #E9 1544388405822 #E9

RESUMEN

Por potencia contratada		#J1 27,98 #J1 €
Por energía consumida		#J2 133,02 #J2 €
Impuesto electricidad		#N2 8,23 #N2 €
Alquiler equipos medida y control		#M4 3,02 #M4 €
IVA	#N6 7 #N6% s:#N7 169,23 #N7 €	#N8 11,85 #N8 €
	#N4 10 #N4% s:#M4 3,02 #M4 €	#N5 0,30 #N5 €
TOTAL IMPORTE FACTURA:		#J5 184,40 #J5 €

Remite: #C1 MORGAN STANLEY CAPITAL GROUP ESPAÑA, S.L. #C1 Apartado de Correos XXXXX [illegible] MADRID #C1

#A1 Celeste Zetina Quesada #A1

#A3 Calle Hospital #A3

#A4 25577 #A4 #A5 Lladorre #A5

Fig. 2. A fragment of an invoice with annotations of the IDSEM dataset. The annotations allow labeling the relevant information of the invoices. The code of the labels surrounds the value in the invoice.

Table 1 shows the file resulting from converting the invoice in Fig. 1 to text format. It presents an extract from the file in Spacy [17] format. Each invoice is described by the text extracted from the PDF along with the entities identified by their start and end positions in the text. Additionally, in Table 1 we provide a list of the various labels identified in the invoice.

We have selected a subset of 5.000 invoices corresponding to five different marketers. A total of 4.000 invoices are used for training and 1.000 for validation. We convert all the invoices from PDF to text and identify the position of the labels in the text. We use the `pdf2text` application for extracting the text from the PDF file. This extraction is done line by line, sometimes causing label information to span multiple lines or overlap with multiple tokens on the same line. We detect the start and end positions of each label using the labels annotations identified in the text, i.e. `(354, 360, 'J5')` for the label 'J5', total price. With this information, we build the entity section of the Spacy file. In Fig. 3, we show the pipeline for converting the IDSEM invoices to Spacy format.

Table 1. An electricity invoice in text format corresponding to the bill shown in Fig. 1. This is the format of the SpaCy library, with the position of entities and their codes at the end. At the bottom, we enumerate the most relevant labels present in the invoice.

```
[('DATOS DE LA FACTURA DE ELECTRICIDAD\n'

'IMPORTE FACTURA: 184,40 €\n'

'N° Factura: L9563906771 emitida el 13 de enero de
2010\n'

'Periodo de consumo: 12 de noviembre de 2009 a #F5
11 de enero de\n'

'Fecha de cargo: 16 de enero de 2010\n'
'Referencia del contrato de suministro:
1544388405822\n'

...

'RESUMEN MADRID #C5\n'

'Por potencia contratada 27,98 €'

'Celeste Zetina Quesada\n'

'Por energía consumida 133,02 €\n'

' 8,23 €\n'

'Calle Hospital\n'

'Impuesto electricidad\n'

'Alquiler equipos medida y control 3,02 € 25577'

'Lladorre\n'

'IVA 11,85 €\n'

' 7 % s/169,23 €\n'

' 0,30 €\n'

' 10 % s/3,02 €\n'

'TOTAL IMPORTE FACTURA: 184,40 €\n'

'\x0c',

{'entities': [(354, 360, 'J5'), (451, 462, 'F1'),

    (474, 493, 'F3'), (590, 613, 'F4'),(1111,
   1130,'G3'), (1246, 1259, 'E9'), (1657, 1662,
  'J1'), (1670, 1692, 'A1'), (1749, 1755, 'J2'),
  (1815, 1819, 'N2'), (1900, 1914, 'A3'), (2008,
  2012, 'M4'), (2021, 2026, 'A4'), (2027, 2035,
  'A5'), (2093, 2098, 'N8'), (2130, 2131, 'N6'),
  (2135, 2141, 'N7'), (2202, 2206, 'N5'), (2240,
     2242, 'N4'), (2246, 2250, 'M4'), (2310,
                 2316,'J5')]})]
```

A1/B1, Customer's name	
A3/B3, Customer's address	N6, Reduced tax rate
A4/B4, Postal code	N8, Tax price
A5/B5, Customer's city	E9, Reference supply contract
J1, Electricity power price	F1, Invoice number
J2, Energy consumed price	F2, Invoice reference
J5, Total price	F3, Invoice release date
M4, Equipment rental price	F4, Start billing date
N2, Electricity tax price	F5, End billing date
N4, Normal tax rate	G3, Payment date
N5, Reduced tax price	

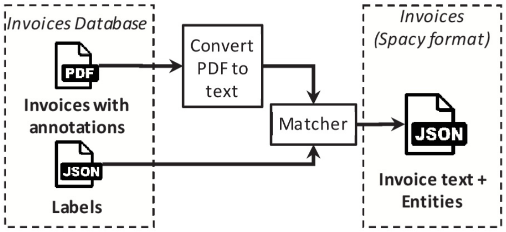


Fig 3. Pipeline for converting the IDSEM invoice to Spacy format, text + entities. In Fig. 1, we have an example of the input data and Table 1, an example of output data.

3.2. Configuration of the NER Model

Named Entity Recognition is a key task in NLP that involves identifying and classifying entities within the text. The NER pipeline typically consists of several steps, including tokenization, part-of-speech tagging, and entity recognition.

The Transformer architecture is based on the self-attention mechanism, which allows the network to attend to different parts of the input sequence with varying degrees of importance. Its architecture consists of an encoder and a decoder, which are connected by multi-headed self-attention and feedforward layers.

We design a system that automatically extracts the relevant information using a NER process. In our method, we choose a Transformer model for training the NER. The system is implemented with the SpaCy [17] library, which is a standard library for NLP in Python.

Fig. 4 shows the pipeline for the training process, which includes the NER component and the input data. The named entity recognizer takes electricity invoices and predicts sentences and their labels in context. The training data includes the text of invoices, the positions of the entities they contain, and the entity label codes. It is necessary to have another entity that do not correspond to any label.

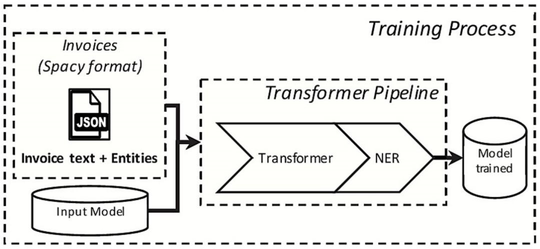


Fig. 4. Transformer pipeline for the training process.

The training data helps the model to specify the information that we want it to predict. Since the model can detect multiple values for the same entity, we fuse the information by checking the format of the label and some statistics of the extracted data. Our trained model can recognize entities in similar contexts, independently of the format employed in each bill and the organization of the entities in the document.

4. Results

To evaluate our model's accuracy, we compare its predictions with invoices not used during training. We calculate the accuracy of our model with standard performance metrics [11], such as Precision, Recall and F1-score. The overall accuracy for all the labels and invoices in the test set is 99.3 %.

Table 2 shows the results for many principal labels, like the customer's name (A1), the total price (J5), the equipment rental price (M4), or the taxes (N2, N5, N8). The precision, recall, and F1-score metrics are presented for each label. There are many labels that obtain an F1-score of 100 %. The F1-score of the N2 label, Electricity tax price, was the lowest one with 41 %. Its recall is low, which indicates many false negatives in this case. These results show that the proposed method achieves high accuracy for most of the labels.

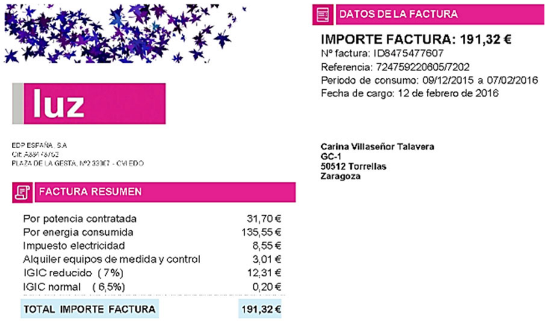
Table 2. Precision, Recall and F1-score obtained for some labels of the invoice of Fig. 1.

Label	Precision	Recall	F1-score
A1/B1, Customer's name	100 %	100 %	100 %
A3/B3, Customer's address	100 %	100 %	100 %
A4/B4, Postal code	100 %	100 %	100 %
A5/B5, Customer's city	100 %	100 %	100 %
C1, Marketer's name	32 %	100 %	48 %
C5, Marketer's city	99 %	99 %	99 %
CD, Customer's support ph.	58 %	100 %	74 %
J1, Electricity power price	94 %	100 %	97 %
J2, Energy consumed price	100 %	100 %	100 %
J5, Total price	100 %	100 %	100 %
M4, Equipment rental price	100 %	100 %	100 %
N2, Electricity tax price	100 %	26 %	41 %
N4, Normal tax rate	100 %	100 %	100 %
N5, Reduced tax price	100 %	100 %	100 %
N6, Reduced tax rate	100 %	100 %	100 %
N8, Tax price	77 %	100 %	87 %
F1, Invoice number	100 %	100 %	100 %
F3, Invoice release date	100 %	100 %	100 %
F4, Start billing date	100 %	100 %	100 %
F5, End billing date	100 %	100 %	100 %

Fig. 9 displays two confusion matrices corresponding to the labels detailed in Table 2. The diagonal values in the matrices indicate the number of correctly identified labels, while the off-diagonal values represent the misclassified labels. The matrices demonstrate high accuracy of the proposed method. The first matrix shows the labels with the highest value of the "Precision" while the second matrix shows the less precise labels.

To verify the correct functioning of the proposed method, Fig. 5 shows an example of an electricity bill used in the tests.

The visualizations produced by `displacy` are informative and provided valuable insights into the model's performance. These visualizations allowed us to identify the entity types that the model performed well on and those that it struggled with. For instance, the model has high precision and recall for identifying the marketer and customer information.



luz

EDP ESPAÑA, S.A
Cif: A33473752
PLAZA DE LA GESTA, Nº2 33007 - OVIEDO

FACTURA RESUMEN

Por potencia contratada	31,70 €
Por energia consumida	135,55 €
Impuesto electricidad	8,55 €
Alquiler equipos de medida y control	3,01 €
IGIC reducido (7%)	12,31 €
IGIC normal (6,5%)	0,20 €
TOTAL IMPORTE FACTURA	191,32 €

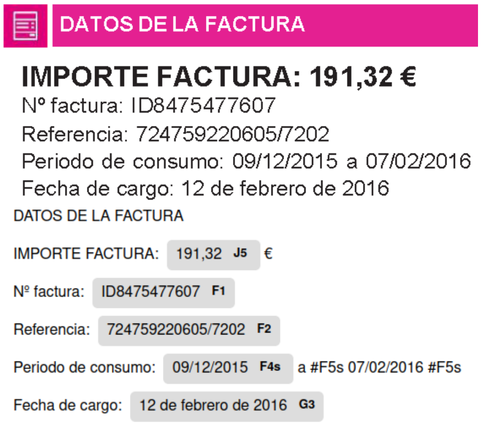
DATOS DE LA FACTURA

IMPORTE FACTURA: 191,32 €
Nº factura: ID8475477607
Referencia: 724759220605/7202
Periodo de consumo: 09/12/2015 a 07/02/2016
Fecha de cargo: 12 de febrero de 2016

Carina Villaseñor Talavera
GC-1
50512 Torrellas
Zaragoza

Fig. 5. A fragment of an invoice of the IDSEM dataset.

Figs. 6, 7 and 8, depict some fragments of the invoice shown in Fig. 5. They highlight named entities and their labels in a text. The figures show that the proposed method successfully identified all the relevant information from the invoice, demonstrating the effectiveness of the proposed method.



DATOS DE LA FACTURA

IMPORTE FACTURA: 191,32 €
Nº factura: ID8475477607
Referencia: 724759220605/7202
Periodo de consumo: 09/12/2015 a 07/02/2016
Fecha de cargo: 12 de febrero de 2016

DATOS DE LA FACTURA

IMPORTE FACTURA: 191,32 J5 €

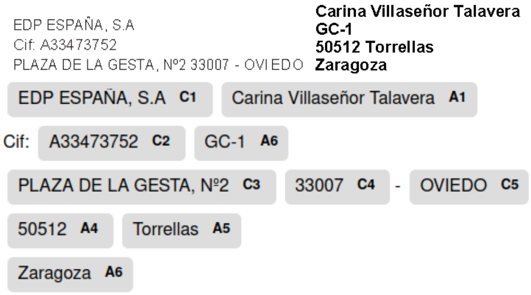
Nº factura: ID8475477607 F1

Referencia: 724759220605/7202 F2

Periodo de consumo: 09/12/2015 F4s a #F5s 07/02/2016 #F5s

Fecha de cargo: 12 de febrero de 2016 G3

Fig. 6. Fragment of the summary of the invoice shown in Fig. 5 and the named entities identified by our method.



EDP ESPAÑA, S.A
Cif: A33473752
PLAZA DE LA GESTA, Nº2 33007 - OVIEDO

Carina Villaseñor Talavera
GC-1
50512 Torrellas
Zaragoza

EDP ESPAÑA, S.A C1 Carina Villaseñor Talavera A1

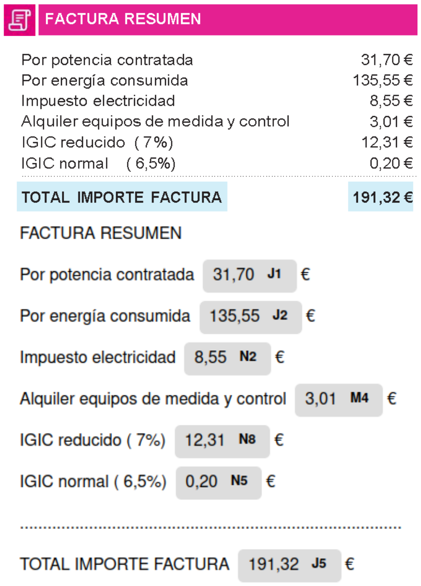
Cif: A33473752 C2 GC-1 A6

PLAZA DE LA GESTA, Nº2 C3 33007 C4 - OVIEDO C5

50512 A4 Torrellas A5

Zaragoza A6

Fig 7. Fragment of the electricity company and the customer of the invoice shown in Fig. 5 and the named entities identified by our method.



FACTURA RESUMEN

Por potencia contratada	31,70 €
Por energía consumida	135,55 €
Impuesto electricidad	8,55 €
Alquiler equipos de medida y control	3,01 €
IGIC reducido (7%)	12,31 €
IGIC normal (6,5%)	0,20 €
TOTAL IMPORTE FACTURA	191,32 €

FACTURA RESUMEN

Por potencia contratada 31,70 J1 €

Por energía consumida 135,55 J2 €

Impuesto electricidad 8,55 N2 €

Alquiler equipos de medida y control 3,01 M4 €

IGIC reducido (7%) 12,31 N8 €

IGIC normal (6,5%) 0,20 N5 €

TOTAL IMPORTE FACTURA 191,32 J5 €

Fig 8. Fragment of the breakdown of the invoice shown in Fig. 5 and the named entities identified by our method.

5. Conclusion

We proposed a method for automatically extracting information from electricity invoices. This method relied on text data and is based on an NLP pipeline to process the information. The main module of the system is based on a NER process, supported by a Transformer neural network, that allows detecting the named entities. The results showed that the proposed system can extract the main labels with high accuracy. The visualizations shown allowed us to gain a deeper insight on the model's strengths and weaknesses. In future works, we will explore other neural network architectures to integrate visual information with text data.

Acknowledgements

This work was partly supported by the Servicio Canario de Salud, Gobierno de Canarias, under Contract F2022/03.

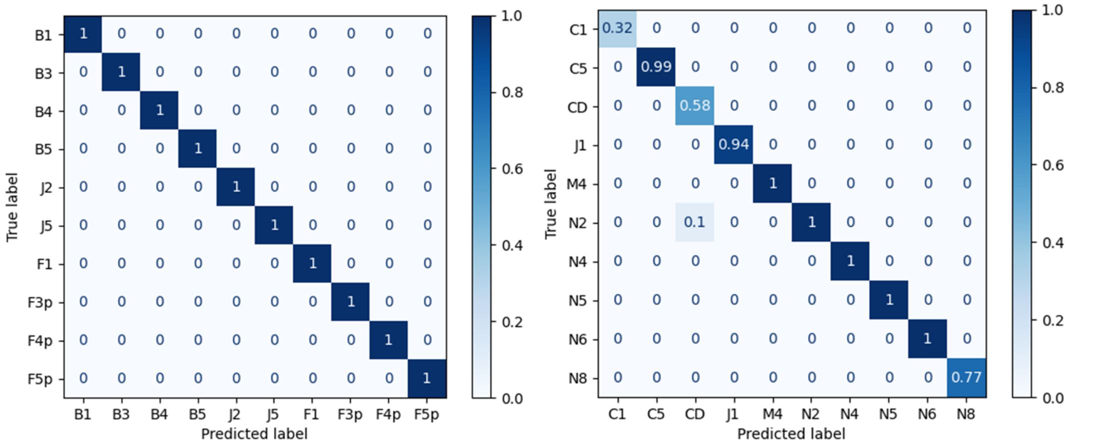


Fig. 9. Confusion matrices of the labels shown in Table 2.

References

[1]. J. Chavat, S. Nesmachnow, J. Graneri, G. Alvez, ECD-UY, detailed household electricity consumption dataset of Uruguay, *Scientific Data*, Vol. 9, 2022, 21.

[2]. V. P. d'Andecy, E. Hartmann, M. Rusiñol, Field extraction by hybrid incremental and a-priori structural templates, in *Proceedings of the 13th IAPR International Workshop on Document Analysis Systems (DAS'18)*, 2018, pp. 251-256.

[3]. A. R. Dengel, B. Klein, Smartfix: A requirements-driven system for document analysis and understanding, in Document Analysis Systems V (D. Lopresti, J. Hu, R. Kashi, Eds.), *Springer*, 2002, pp. 433-444.

[4]. D. Esser, D. Schuster, K. Muthmann, M. Berger, A. Schill, Automatic indexing of scanned documents: A layout-based approach, *Proceedings of SPIE*, Vol. 8297, 2012, pp. 118-125.

[5]. G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, *arXiv Preprint*, 2016, arXiv:1603.01360.

[6]. X. Liu, F. Gao, Q. Zhang, H. Zhao, Graph convolution for multimodal information extraction from visually rich documents, *arXiv Preprint*, 2019, arXiv:1903.11279.

[7]. S. Makonin, B. Ellert, I. Bajic, F. Popowich, Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014, *Scientific Data*, Vol. 3, 2016, 160037.

[8]. I. Muslea, Extraction patterns for information extraction tasks: A survey, in *Proceedings of the AAAI Workshop on Machine Learning for Information Extraction*, Vol. 2, 1999.

[9]. R. B. Palm, F. Laws, O. Winther, Attend, copy, parse end-to-end information extraction from documents, in *Proceedings of the International Conference on*

Document Analysis and Recognition (ICDAR'19), 2019, pp. 329-336.

[10]. R. B. Palm, O. Winther, F. Laws, Cloudscan – A configuration-free invoice analysis system using recurrent neural networks, in *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR'17)*, Vol. 1, 2017, pp. 406-413.

[11]. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, Vol. 12, 2011, pp. 2825-2830.

[12]. M. Rusiñol, T. Benkhelfallah, V. P. d'Andecy, Field extraction from administrative documents by incremental structural templates, in *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR'13)*, 2013, pp. 1100-1104.

[13]. C. Sage, A. Aussem, H. Elghazel, V. Eglin, J. Espinas, Recurrent neural network approach for table field extraction in business documents, in *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'19)*, 2019, pp. 1308-1313.

[14]. J. Sánchez, A. Salgado, A. García, N. Monzón, IDSEM, an invoices database of the Spanish electricity market, *Scientific Data*, Vol. 9, 2022, 786.

[15]. D. Schuster, K. Muthmann, D. Esser, A. Schill, M. Berger, C. Weidling, K. Aliyev, A. Hofmeier, Intellix-end-user trained information extraction for document archiving, in *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR'13),* 2013, pp. 101-105.

[16]. C. Shin, E. Lee, J. Han, J. Yim, W. Rhee, H. Lee, The ENERTALK dataset, 15 Hz electricity consumption data from 22 houses in Korea, *Scientific Data*, Vol. 6, 2019, 193.

[17]. SpaCy: Industrial Strength Natural Language Processing in Python, https://spacy.io/

[18]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in Advances in Neural Information Processing Systems, Vol. 30, *Curran Associates, Inc*., 2017.

[19]. Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, Layoutlm: Pre-training of text and layout for document image understanding, in *Proceedings of the Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'20)*, New York, NY, USA, 2020, pp. 1192-1200.

[20]. X. Zhao, E. Niu, Z. Wu, X. Wang, Cutie: Learning to understand documents with convolutional universal text information extractor, *arXiv Preprint*, 2019, arXiv:1903.12363.