

Spatial Dependency Parsing for Semi-Structured Document Information Extraction

Wonseok Hwang Jinyeong Yim Seunghyun Park Sohee Yang Minjoon Seo

Clova AI, NAVER Corp.

{wonseok.hwang, jinyeong.yim, seung.park, sh.yang, minjoon.seo}
@navercorp.com

Abstract

Information Extraction (IE) for semi-structured document images is often approached as a sequence tagging problem by classifying each recognized input token into one of the IOB (Inside, Outside, and Beginning) categories. However, such problem setup has two inherent limitations that (1) it cannot easily handle complex spatial relationships and (2) it is not suitable for highly structured information, which are nevertheless frequently observed in real-world document images. To tackle these issues, we first formulate the IE task as *spatial dependency parsing* problem that focuses on the relationship among text tokens in the documents. Under this setup, we then propose SPADE Δ (SPAtial DEpendency parser) that models highly complex spatial relationships and an arbitrary number of information layers in the documents in an end-to-end manner. We evaluate it on various kinds of documents such as receipts, name cards, forms, and invoices, and show that it achieves a similar or better performance compared to strong baselines including BERT-based IOB tagger.

1 Introduction

Document information extraction (IE) is the task of mapping each document to a structured form that is consistent with the target ontology (e.g., database schema), which has become an increasingly important task in both research community and industry. In this paper, we are particularly interested in information extraction from real-world, semi-structured document images, such as invoices, receipts, and name cards, where we assume Optical Character Recognition (OCR, i.e. detecting the locations of the text tokens if the input is an image) has been already applied. Previous approaches for semi-structured document IE often assume as if

the input is a one-dimensional sequence and formulate the task as an IOB (Inside Outside Beginning) tagging problem. In this setup, the tokens in the document (either obtained through an OCR engine or trivially parsed from a web page or pdf) are first serialized, and then an independent tagging model classifies each of the flattened lists into one of the pre-defined IOB categories (Ramshaw and Marcus, 1995; Palm et al., 2017). While effective for relatively simple documents, their broader application in the real world is still challenging because (1) semi-structured documents often exhibit a complex layout where the serialization algorithm is non-trivial, and (2) sequence tagging is inherently not effective for encoding multi-layer hierarchical information such as the menu tree in receipts (Fig. 1c).

To overcome these limitations, we propose SPADE Δ (SPAtial DEpendency parser), an end-to-end, serializer-free model that is capable of extracting hierarchical information from complex documents. Rather than explicitly dividing the original problem into two independent subtasks of serialization and tagging, our model tackles the problem in an end-to-end manner by creating a directed relation graph of the tokens in the document (Fig. 1). In contrast to traditional dependency parsing, which parses the dependency structure in purely (one-dimensional) linguistic space, our approach leverages both linguistic and (two-dimensional) *spatial* information to parse the dependency.

We evaluate SPADE Δ on eight document IE datasets created from real-world document images, including invoices, name cards, forms, and receipts, with the varying complexity of information structure. In all of the datasets, our model shows a similar or better accuracy than strong baselines including BERT-based IOB taggers, and particularly outstands in documents with complex layouts (Table 3). These results demonstrate the effectiveness

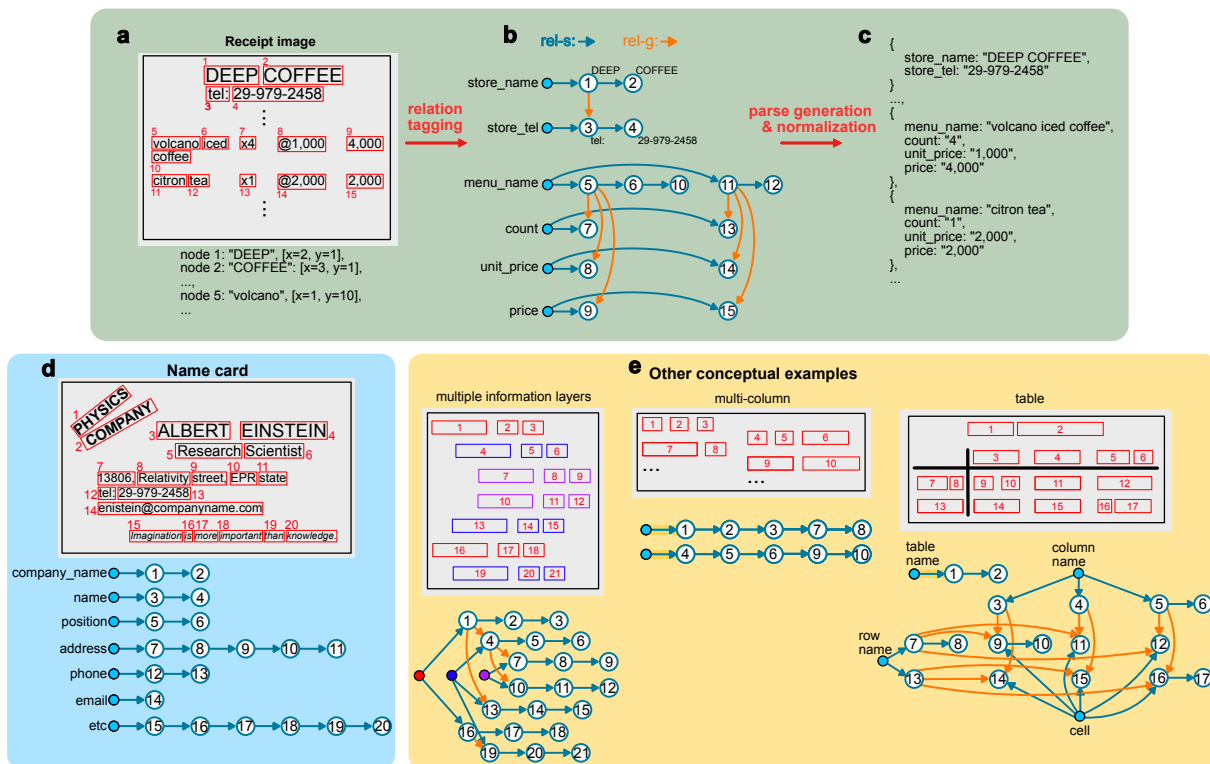


Figure 1: The illustration of spatial dependency parsing problem. Receipt parsing is explained in detail with three subfigures: (a) first, text tokens and their coordinates are extracted from OCR; (b) next, the relations between tokens are classified into two types: *rel-s* for serialization and information type (field) classification, and *rel-g* for inter-grouping between fields (the numbers inside of circles in (b) indicates the box numbers in (a)); (c) the final parse is generated by decoding the graph. (d) A sample name card and its spatial dependency parse. (e) Other conceptual examples showing the versatility of the spatial dependency parsing approach for document IE.

of our end-to-end, graph-based paradigm over the existing sequential tagging approaches.

In short, our contributions are threefold. (1) We present a novel view that information extraction for semi-structured documents can be formulated as a dependency parsing problem in two-dimensional space. (2) We propose SPADE \diamond for spatial dependency parsing, which is capable of efficiently constructing a directed semantic graph of text tokens in semi-structured documents.¹ (3) SPADE \diamond achieves a similar or better accuracy than the previous state of the art or strong BERT-based baselines in eight document IE datasets.

2 Related Work

The recent surge of interest in automatic information extraction from semi-structured documents are well reflected in their increased number of publication record from both research community and industry (Katti et al., 2018; Qian et al., 2019; Liu et al., 2019; Zhao et al., 2019; Denk and Reisswig, 2019; Hwang et al., 2019; Park et al., 2019; Xu

et al., 2019; Jaume et al., 2019; Zhong et al., 2019; Rausch et al., 2019; Yu et al., 2020; Wei et al., 2020; Majumder et al., 2020; Lockard et al., 2020; Ganczarek et al., 2020; Lin et al., 2020; Xu et al., 2020; Powalski et al., 2021; Wang et al., 2021; Hong et al., 2021; Hwang et al., 2021). Below, we summarize some of closely related works published before the major development of SPADE \diamond .

Serialized IE Previous semi-structured document information extraction (IE) methods often require the input text boxes (obtained from OCR) to be serialized into a single flat sequence. Hwang et al. (2019) and Denk and Reisswig (2019) combine a manually engineered text serializer that turn the OCR text boxes into a sequence and a Transformer-based encoder, BERT (Devlin et al., 2018), that performs IOB tagging on the sequence or semantic segmentation from images. In contrast to SPADE \diamond , these models rely on the serialization of the tokens and thus it is difficult to flexibly apply them to documents with complex layouts such as multi-column or distorted documents. Xu et al. (2019) propose LayoutLM that jointly embeds the

¹<https://github.com/clovaai/spade>

image segments, text tokens, and positions of the tokens in an image to make a pretrained model for document understanding. However, LayoutLM still requires a careful serialization of the tokens as it relies on the position embeddings of BERT. Also, it is only evaluated on classification for the downstream task.

Serializer-free IE Existing serializer-free methods mostly extract flat key-value pairs, as they still formulate the task as tagging the text tokens. They fundamentally differ from SPADE \diamond which generates a structured output that captures full information hierarchy represented in the document. Chargrid (Katti et al., 2018) performs semantic segmentation on invoice images to extract target key-value pairs. Although Chargrid uses additional “bounding boxes” for inter-grouping of certain fields, the application to the documents that have more than two information hierarchy levels is non-trivial. Also, when fields that belong to the same group are remotely located, the bounding boxes may need to be modified to have a more complex geometrical shape to avoid overlap between the boxes.

Graph-based IE Liu et al. (2019); Qian et al. (2019); Wei et al. (2020); Yu et al. (2020) utilize a graph convolution network to contextualize the tokens in a document and a bidirectional LSTM with CRF to predict the IOB tags. However, the range of possible parse generations is limited as IOB tagging can be performed only within each OCR bounding box, ignoring inter-box relationship. On the contrary, SPADE \diamond predicts both the intra-box relationship and the inter-box relationship by constructing a dependency graph among the tokens.

Lockard et al. (2019, 2020) also utilize a graph to extract semantic relation from semi-structured web-page. The graph is constructed based on rules from “structured html DOM” and mainly used for information encoding. On the other hand SPADE \diamond accepts “unstructured text distributed in 2D” and generates graphs as the result of decoding (in a data-driven way).

Dependency parsing Dependency parsing is the task of obtaining the syntactic or semantic structure of a sentence by defining the relationships between the words in the sentence (Zettlemoyer and Collins, 2012; Peng et al., 2017; Dozat and Manning, 2018). The relations are often expressed as directed, labeled arcs. In our work, we view the problem of

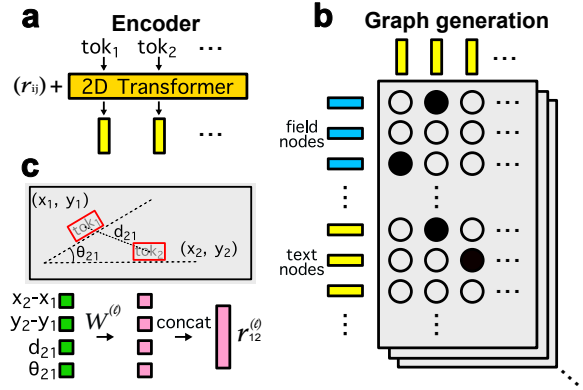


Figure 2: The illustration of SPADE \diamond . (a) Spatial text encoder contextualizes tokens using their relative spatial vectors $\{r_{ij}\}$ (Eq. 1). (b) The dependency graph is inferred by mapping the vector representations of each pair of tokens into a scalar value. The field embeddings are blue-colored. (c) At each encoder layer l , r_{ij}^l is prepared by concatenating four embedding vectors: relative coordinates, distance, and angles embeddings. W^l stands for a linear projection.

information extraction for semi-structured documents as a *spatial* dependency parsing task such that two-dimensional spatial information is mainly considered. This setup enables SPADE \diamond to flexibly handle documents with complicated layouts while representing the full information hierarchy.

3 Problem definition

In this section, we first describe the task of information extraction for semi-structured documents, and we briefly discuss how the task was approached in the past as a sequence tagging problem. Then we formulate it as a spatial dependency parsing problem. In Section 4, we show how we design our model for the newly formulated problem.

3.1 Semi-structured document IE

Document IE is often defined as the extraction of structured information (e.g. key-value pairs) in documents. For semi-structured documents, the task becomes more challenging, mainly due to two factors: (1) complex spatial layout and (2) hierarchical information structure. In the simplest case, both of the two factors are minimally present, where the text is strictly a linear sequence, and the desired output is simply a list of fields, similar to Named Entity Recognition (NER) task. However, the problem becomes more difficult when at least one of the factors is significant. In name cards, spatial relationship can be tricky; Fig. 1d shows an example where a naive left-to-right serialization would fail

because the `company_name` (“Physics Company”) is tilted. In receipts, their hierarchical information structure complicates the problem. For example, in Fig. 1a), words “volcano” (box 5), “iced” (box 6), and “coffee” (box 10) together form a single field `menu_name`, and the field constitutes another group in the second hierarchical layer with the `count` field (box7), `unit_price` field (box 8), and `price` field (box 9). Other conceptual examples are shown in Fig. 1e); documents that have triple information layers (left), multiple columns (middle), and a table (right).

3.2 Previous formulation: Sequence tagging

As mentioned, IOB sequence tagging is appropriate for document IE when the layout and the information structure are simple (Ramshaw and Marcus, 1995; Lample et al., 2016; Chiu and Nichols, 2016; Ma and Hovy, 2020). When one of the factors is present, however, one has to adopt an ad-hoc solution to detour the inherent limitation of IOB.

In the case of complex spatial relationship (e.g., name card), an advanced, dedicated serialization method can be considered. However, it may require layout-specific manual engineering, which becomes more difficult for documents such as name cards that exhibit diverse layouts.

In the case of complex information structure (e.g., receipt), one can consider augmenting each IOB tag with higher-layer information. For instance, in a typical IOB setting, the `menu_name` field will require two tags, namely `menu_name_B` and `menu_name_I`. To model the second layer information (inter-grouping of fields), `menu_name_B` can be augmented into two, namely `B2_menu_name_B`, `I2_menu_name_B`, where `B2` and `I2` indicate the beginning and the inside of the hierarchy’s second layer. While effective for some applications, this method would not generalize well to an arbitrary depth as it requires more tags for each additional layer.

3.3 Our formulation: Spatial dependency parsing

To better model spatial relationship and hierarchical information structure in semi-structured documents, we formulate the IE problem as “spatial dependency parsing” task by constructing a dependency graph with tokens and fields as the graph nodes (node per token and field type). This is demonstrated in Fig. 1, where empty blue circles are text nodes, and filled blue circles are field

nodes.

Although the spatial layout of semi-structured documents is diverse, it can be considered as the realization of mainly two abstract properties between each pair of nodes, (1) `rel-s` for the ordering and grouping of tokens belonging to the same information category (blue arrows in Fig. 1b), and (2) `rel-g` for the inter-group relation between grouped tokens or groups (orange arrows in the same figure). Connecting a field node to a text node indicates that the text is classified into the field. For example, “volcano iced coffee” in Fig. 1a) is classified as a menu name by being attached to the `menu_name` field node with blue arrows, and it is connected with “x4”, “@1,000”, and “4,000” with orange arrows to indicate the hierarchical information among the groups. The dependency graphs of name cards and other conceptual examples are also shown in Fig. 1d and e.

4 Model

To perform the spatial dependency parsing task introduced in the previous section in an end-to-end fashion, we propose SPADE Δ that consists of (1) spatial text encoder, (2) graph generator, and (3) graph decoder. Spatial text encoder and graph generator are trained jointly. Graph decoder is a deterministic function (without trainable parameters) that maps the graph to a valid parse of the output structure.

4.1 Spatial text encoder

Spatial text encoder is based on 2D Transformer architecture. Unlike the original Transformer (Vaswani et al., 2017), there is no order among the input tokens, making the model invariant under the permutation of the input tokens. Inspired by Transformer XL (Dai et al., 2019), the attention weights (between each key and query vector) is computed by

$$q_i^T k_j + q_i^T r_{ij} + (b_i^{key})^T k_j + (b_i^{rel})^T r_{ij} \quad (1)$$

where q_i is the query vector of the i -th input token, k_j is the key vector of the j -th input token, r_{ij} is the relative spatial vector of the j -th token with respect to the i -th token, and $b_i^{key/rel}$ is a bias vector. In (original) Transformer, only the first term of Equation 1 is used.

The relative spatial vector r_{ij} is constructed as follows (Fig. 2c). First, the relative coordinates be-

tween each pair of tokens are computed.² Next, the coordinates are quantized into integers and embedded using \sin and \cos functions (Vaswani et al., 2017). The physical distance and the relative angle between each pair of the tokens are also embedded in a similar way. Finally, the four embedding vectors are linearly projected (with a trainable projection matrix) and concatenated at each encoder layer.

4.2 Graph generator

As discussed in Section 3.3 and shown in Fig. 1, every token corresponds to a node and each pair of the nodes forms one of the two relations (or no relation): (1) `rel-s` for serializing tokens within the same field, and (2) `rel-g` for inter-grouping between fields. The dependency graph can be represented by using a binary matrix $M^{(r)}$ for each relation type r (Fig. 2b) where $M_{ij}^{(r)} = 1$ if there exists a directed edge from the i -th token to the j -th token and 0 otherwise. Each $M^{(r)}$ consists of $n_{\text{field}} + n_{\text{text}}$ number of rows and n_{text} number of columns where n_{field} and n_{text} represent the number of field types and the number of tokens, respectively. The graph generation task now becomes predicting the binary matrix.

We obtain $M^{(r)}$ as follows. The probability that there exists a directed edge $i \xrightarrow{r} j$ is computed by

$$\begin{aligned} h_i^{(r)} &= \begin{cases} u_i^{(\text{field})}, & \text{for } i \leq n_{\text{field}} \\ \mathcal{W}_h^{(r)} v_i & \text{otherwise} \end{cases} \\ d_j^{(r)} &= \mathcal{W}_d^{(r)} v_j \\ s_{0,ij}^{(r)} &= (h_i^{(r)})^T \mathcal{W}_0^{(r)} d_j^{(r)} \\ s_{1,ij}^{(r)} &= (h_i^{(r)})^T \mathcal{W}_1^{(r)} d_j^{(r)} \\ p_{ij}^{(r)} &= \frac{\exp(s_{1,ij}^{(r)})}{\exp(s_{0,ij}^{(r)}) + \exp(s_{1,ij}^{(r)})}, \end{aligned} \quad (2)$$

where $u_i^{(\text{field})}$ represents the trainable embedding vector of the i -th field type node (filled blue circles in Fig. 1), $\{v_i\}$ is a set of vectors of contextualized tokens from the encoder, \mathcal{W} stands for affine transformation, h is the embedding vector of the head token, and d is that of the dependent token.

²For example, if “token1” is at $(x_1 = 1, y_1 = 10)$ and “token2” is at $(x_2 = 3, y_2 = 4)$, the relative coordinate of “token2” with respect to “token1” is $(x'_2, y'_2) = (3-1, 4-10) = (2, -6)$.

$M_{ij}^{(r)}$ is obtained by binarizing $p_{ij}^{(r)}$ as follows.

$$M_{ij}^{(r)}(p_{ij}^{(r)}) = \begin{cases} 1 & \text{for } (r=s, i \text{ is field type node, } p_{ij}^{(r)} \geq p_{th}) \\ 1 & \text{for } (r=s, i \text{ is text node, } p_{ij}^{(r)} \geq p_{th}, \\ & j = \arg \max_k p_{ik}^{(r)}) \\ 1, & \text{for } (r=g, i \text{ is text node, } p_{ij}^{(r)} \geq p_{th}) \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

The recall rate of edges can be controlled by varying the threshold value p_{th} . Here, we set $p_{th} = 0.5$.

Tail collision avoidance algorithm Each node in spatial dependency graphs has a single incoming edge per relation except some special documents such as table (Fig. 1e). Based on this property, we apply the following simple yet powerful tail collision avoidance algorithm: (1) at each tail node having multiple incoming edges, all edges are trimmed except the one with the highest linking probability; (2) at each head node of the trimmed edges, the new tail node is found by drawing the next probable edge whose probability is larger than p_{th} and belongs to the top three; (3) go back to Step 1 and repeat the routine until the process becomes self-consistent or the max iteration limit is reached (set to 20 in this paper). The algorithm prevents loops and token redundancy in parses.

4.3 Graph decoder

We decode the generated graph into the final parse through the following three stages: (1) SEEDING, (2) SERIALIZATION, and (3) GROUPING (Table 1). In SEEDING, field type nodes (filled circles in Fig. 1) are linked to multiple text nodes (seeds) by `rel-s`. In SERIALIZATION, each seed node found in the previous stage generates a directed edge (`rel-s`) to the next text node (i.e. serialization) recursively until there is no further node to be linked. Finally, in GROUPING, the serialized texts are grouped iteratively, constructing information layers from the top to the bottom. The total number of iterations is equal to “the number of information layers-1”. To group texts using directed edges, we define a special representative field for each information layer. Then, the first token of the representative field generates directed edges to the first token of other fields that belong to the same group using `rel-g` (for example, `menu_name` (“volcano iced coffee”) in Fig. 1a) generates directed edges to other member fields (`count` (“x4”), `unit_price` (“@1,000”) and `price` (“4,000”)).

The process generates an *arborescence*³ for each field (re1-s) and group (re1-g). The resulting set of graphs has a one-to-one correspondence with the parse through detokenization. The use of beam search in SERIALIZATION does not introduce noticeable difference in re1-s probably due to the short decoding length of the graph (mostly less than 30). The development of a more advanced decoding algorithm that generates globally optimal multiple arborescences remains as future work.

Although undirected edges can be employed for the inter-grouping of fields, the use of directed edges has the following merits: (1) an arbitrary depth of information hierarchy can be described without increasing the number of relation types (Fig. 1e) under a unified framework and (2) a parse can be generated in a straightforward manner by iteratively selecting dependent nodes.

Table 1: A formal description of the parse decoding process. s and g stand for re1-s and re1-g respectively.

Action	Input node	Graph at time $t + 1$
INITIALIZATION		$G_{t=0} = \text{empty set}$
SEEDING(μ)	$\mu \in \text{field nodes}$	$G^{(\text{seed})} = \{\mu \xrightarrow{s} j M_{ij}^{(s)} = 1\}$
SERIALIZATION(i)	$i \in G^{(\text{seed})} \cup G_t$	$G_{t+1} = G_t \cup \{i \xrightarrow{s} j M_{ij}^{(s)} = 1\}$
GROUPING(i)	$i \in G^{(\text{seed})} \cup G_t$ i linked to representer fields	$G_{t+1} = G_t \cup \{i \xrightarrow{g} j M_{ij}^{(g)} = 1\}$
MERGE		$G_t = G_t \cup G^{(\text{seed})}$

5 Experimental Setup

5.1 Optical character recognition

To extract the visually embedded texts from an image, we use our in-house OCR system that consists of CRAFT text detector (Baek et al., 2019b) and Comb.best text recognizer (Baek et al., 2019a). The OCR models are finetuned on each of the document IE datasets. The output tokens and their spatial information on the image are used as the inputs to SPADE \diamond .

5.2 Training

We use 12 layers of 2D Transformer encoder (Section 4.1). The parameters are initialized from bert-multilingual (Devlin et al., 2018)⁴. ADAM optimizer (Kingma and Ba, 2015) is used with the following learning rates: 1e-5 for the encoder, 1e-4 for the graph generator, and 2e-5 for

³A directed graph in which, for a vertex u called the root and any other vertex v , there is exactly one directed path from u to v (Excerpted from Wikipedia)

⁴<https://github.com/huggingface/transformers>

S+BERT+IOB₂ and S_{ADV}+BERT+IOB₂. The decay rates are set to $\beta_1 = 0.9, \beta_2 = 0.999$. The batch size is chosen between 4 and 12. SPADE \diamond is trained by using one to eight NVIDIA V100 or P40 GPUs for two to seven days, depending on the tasks. The dev sets are used to pick the best model except FUNSD task in which the model is trained in two steps. First, the 25 examples from training set are sampled and used for a model validation. Next, the model is further trained using entire training set and stopped after 1000 epochs. The training dataset is augmented by randomly rotating the text coordinates by a degree of -10° to $+10^\circ$, (2) by distorting the whole coordinates randomly using a trigonometric function, and (3) by randomly deleting or inserting a single token with 3.3% probability each. Also, 1–2 random tokens from training is attached at the end of the text segments from OCR bounding box with 1.7% probability each. In NAMECARD task, the tokens are not augmented. The identical augmentation algorithm are applied to S+BERT+IOB₂, S_{ADV}+BERT+IOB₂ and SPADE \diamond .

5.3 Evaluation metric

To evaluate the predicted parses that consist of hierarchically organized key-value pairs (e.g. Fig. 3, Fig. 4, 5, 6 in Appendix) we use F_1 score based on exact match. First the group of key-value pairs between predictions and ground truth (gt) are matched based on their string edit distance. Each key-value pairs in the predicted parse is counted as true positive if same key-value pair exists within the corresponding group in gt. Otherwise it is counted as false positive. The unmatched key-value pairs in gt are counted as false negative. The accuracy of dependency parsing is evaluated by computing F_1 of predicted edges. For FUNSD dataset, entity labeling and entity linking scores are computed following the original paper (Jaume et al., 2019). See Appendix A.2 for more details.

5.4 Data statistics

We summarize the data statistics in Table 2, 6. The property of each dataset and their collection process is described in Appendix A.1.

6 Experimental Results

The main focus of SPADE \diamond is to handle the two challenging factors of semi-structured document information extraction—complex spatial relationships and highly structured information—in a gen-

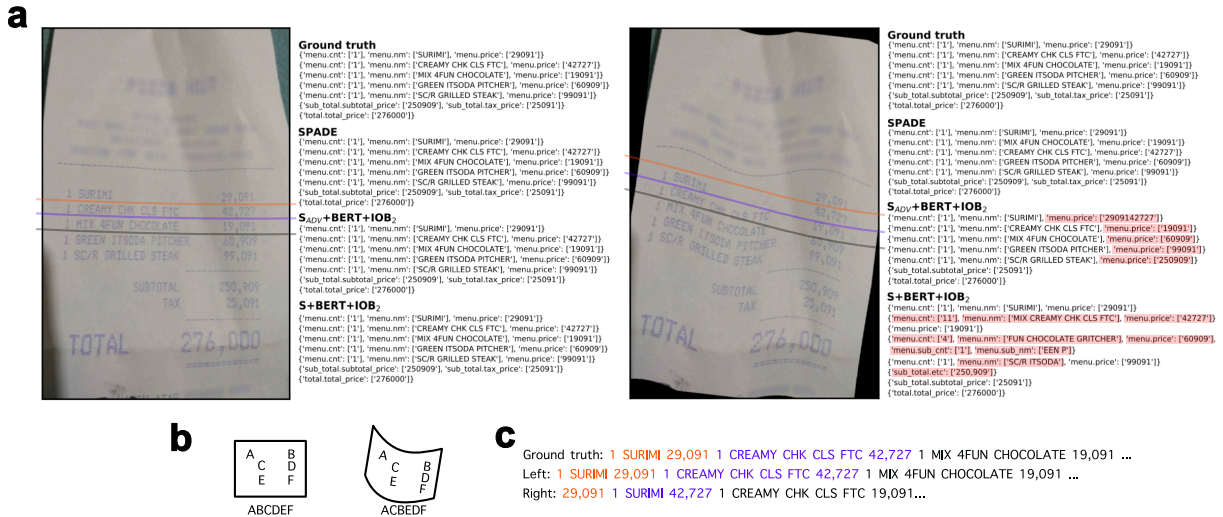


Figure 3: Examples from CORD (left) and CORD++(right) dev sets. (a) Parses are shown in grouped key-value format with the errors in red. (b) The illustration of serialization error. (c) The input tokens serialized by S_{adv} .

Table 2: The dataset properties.

Dataset	Lang.	Abbr.	# of field types	# of examples (train:dev:test)	# of fields	Mean # of text nodes	Depth	Layout complexity
CORD	IDN	co	30	800:100:100	13030	62.3	2	low
CORD+	IDN	co+	"	"	"	62.3	2	high
CORD++	IDN	co++	"	"	"	62.3	2	high
CORD-M	IDN	co-M	"	400:50:50	"	124.6	3	low
RECEIPT-IDN	IDN	RI	50	9508:458:450	209728	209	2	low
NAMECARD	JPN	NC	12	22076:256:100	231528	19.4	1	high
INVOICE	JPN	INV	62	896:79:83	37115	412	2	high
FUNSD ^a	ENG	FU	4	149:50	9743	179	3	high

^a The statistics are from Jaume et al. (2019).

erizable way. We first show that our model can handle hierarchical structure in documents by evaluating the model on two datasets CORD (Park et al., 2019) and RECEIPT-IDN that consist of (Indonesian) receipt images. We then show SPADE Δ can perform well on tasks that require modeling the complex spatial relationship in documents by reporting the performance on name card IE where the spatial layout is more complex than receipts. Then the evaluation on the invoice dataset shows the advantage of SPADE Δ when both of the two challenging factors are simultaneously present. Finally, we show that SPADE Δ can handle even more types of documents by evaluating the model on a form understanding dataset, FUNSD (Jaume et al., 2019). Table 3 summarizes the performance of several baseline models and SPADE Δ in various semi-structured document information extraction tasks.

Handling hierarchical structure in documents

CORD consists of receipt images without creases or warping. SPADE Δ initially achieves 91.5% and 87.4% in F_1 with and without the oracle (ground truth OCR results), respectively (Table 3, 1st row, co). Their dependency parsing score is also shown

Table 3: Parse prediction accuracy. The datasets are referred by their abbreviations in Table 2. ΔF_1 indicates the difference between SPADE Δ (2nd row) and $S_{adv}+BERT+IOB_2$ (4th row).

Model	test (+oracle) [†]				test					
	co	RI	NC	INV	co	co+	co++	RI	NC	INV
SPADE Δ w/o TCA	91.5	92.7	94.0	87.4	87.4	86.1	82.6	88.5	91.1	84.5
SPADE Δ	92.5	93.3	94.3	88.1	88.2	87.4	83.1	89.1	91.6	85.0
S+BERT+IOB ₂	92.4*	93.3*	-	-	90.1	74.0	52.0	88.1	-	-
$S_{adv}+BERT+IOB_2$	92.5*	93.4*	94.4*	84.9*	91.5	85.4	64.8	89.3	90.5	83.1
ΔF_1	0	-0.1	-0.1	+3.2	-1.9	+2.0	+18.3	-0.2	+1.1	+1.9
UB-FLAT	58.1	65.4	100	83.2	-	-	-	-	-	-

[†] The input tokens are recognized by human annotators.

* The input tokens are line-grouped by human annotators.

in Table 7 in Appendix (1st panel, co). To push the performance further, we notice that individual text nodes have a single incoming edge for each relation except in special documents like table (Fig. 1). Using this property, we integrate Tail Collision Avoidance algorithm (TCA) that iteratively trims the tail-sharing-edges and generate new edges until the process becomes self-consistent (Section 4.2). F_1 increases by +1.0% and +0.8% with and without the oracle upon the integration (2nd row, co).

Importance of generating hierarchical structure in receipt IE

In receipt IE task, the inter-grouping of fields is critical due to multiple appearance of same field types such as menu_name and price (Fig. 3a). Without the field grouping, the maximum achievable score is 58.1 F_1 (Table 3, 6th row, UB-FLAT). Generating hierarchical parses from the semi-structured documents is relatively new and thus the direct comparison to previous state-of-the-art methods are not feasible without considerable modification. General confidential issue related to industrial documents and multi-lingual

properties of our task also hinder the comparison. In this regard, we build our own baselines consisting of the manually engineered serializer and BERT-based double IOB taggers (S+BERT+IOB₂⁵).

BERT-tagger The serializer generates pseudo-1D-text from the input tokens distributed in 2D and groups them line-by-line based on their height differences. BERT+IOB₂ predicts the boundary between the fields and between the groups of the fields (see Section 3.2 for the detail). In CORD, S+BERT+IOB₂ shows comparable performance with SPADE \diamond with the oracle (-0.1 F_1) but shows +1.9 F_1 on the test set (2nd and 3rd rows, co). The relatively lower score of SPADE \diamond on the test set may originate from the small size of the training set (800, Table 2) as SPADE \diamond needs to handle the text serialization in a data-driven way. Indeed, when both models are trained using RECEIPT-IDN that consists of 9508 training examples, SPADE \diamond outperforms by +1.0 F_1 on the test set (2nd and 3rd rows, RECEIPT-IDN).

Inflexibility of tagging model in handling complex spatial relationships Next, we prepare CORD+ and CORD++, which are more challenging setups where the images are warped or tilted as often seen in real-world applications (Fig. 3). SPADE \diamond significantly outperforms S+BERT+IOB₂ (+13.4% F_1 in CORD+, +31.1% F_1 in CORD++). This is due to the failure in the serialization in S+BERT+IOB₂ resulting in line-mixing (Fig. 3b, c and Fig. 5, 6 in Appendix). To understand how much improvement can be achieved through further manual engineering, we prepare S_{ADV}+BERT+IOB₂ which is equipped with the advanced serializer where polynomial fitting is employed to group tokens placed on curvy line. The result shows although there is a large improvement in CORD+ and CORD++ task compared to S+BERT+IOB₂, SPADE \diamond still shows the better performance (+2.0% in CORD+, +18.3% in CORD++, 1st and 4th rows). This shows the limitation of a serializer-based method that it cannot be easily generalized to handle document images in wild and the performance can be bottlenecked by the serialization step regardless of how advanced tagging models are. The competent performance of SPADE \diamond on CORD-M, a dataset generated by concatenating two receipt images from CORD into a single image (Fig. 4 in Appendix), further high-

lights the flexibility of SPADE \diamond .

Handling documents having complex layout

We further evaluate SPADE \diamond on name card IE task. Unlike receipts, no inter-grouping between fields is necessary for name card IE. However, name cards often have a complex layout such as non-horizontal alignment of text or multi column even without tilting and warping (Fig. 1d). Our model achieves +1.1% F_1 compared to S_{ADV}+BERT+IOB₂ on the test set (Table 3, NC).

Handling documents having both hierarchical structure and complex layout

To fully explore the capability of SPADE \diamond , we further evaluate the model on invoice IE task. Typical invoices have a hierarchical structure where some fields need to be grouped together, such as `item_name`, `count`, and `price` that correspond to one same item. In addition, invoices also have a relatively complex layout, having multiple tables or columns. SPADE \diamond achieves +1.9 F_1 compared to S_{ADV}+BERT+IOB₂ (Table. 3, INV).

Handling general documents

In order to see if SPADE \diamond can handle more general kinds of documents, we use the FUNSD form understanding dataset (Jaume et al., 2019) where document IE is performed under a more abstract setting by finding general key-value pairs and their inter-grouping (Section A.1.6). The performance is measured on two OCR-independent subtasks (Jaume et al., 2019): (1) “entity-labeling (ELB)” which predicts the information category of the serialized words, and (2) “entity-linking (ELK)” which measures the score for key-value pair link prediction. The evaluation reveals that SPADE \diamond achieves the state of the art on ELK, outperforming the previous baseline by 37.3% F_1 (Table 4, rightmost column). In ELB, SPADE \diamond achieves +11.5% F_1 absolute improvement with respect to BERT-Base Tagger. Both models use BERT-Base as a backbone. Although the F_1 scores of LayoutLM are higher than our model, their contributions are orthogonal to ours since they focus on making a better pretrained model. Also, it cannot perform ELK. We emphasize that SPADE \diamond solves the three subtasks—ELB, ELK, and word serialization—simultaneously, while other tagger models need to use the perfectly serialized input text and solve only entity labeling. The stable performance of SPADE \diamond over randomly rotated documents (ELB-R) or shuffled tokens (ELB-S) supports this highlighting the merit of the serializer-

⁵S stands for the serializer.

free architecture.

Table 4: F_1 scores for two FUNSD subtasks: entity labeling (ELB, ELB-R, and ELB-S) and entity linking (ELK). “Need S” means the input tokens should be serialized. “# of D” indicates the number of documents used for layout pretraining.

Model	Need S # of D	ELB	ELB-R	ELB-S	ELK
Baseline ^a	○ 0	57	-	-	4
BERT-Base Tagger*	○ 0	60.1	43.9 (-16.2)	42.5 (-17.6)	-
BERT-Large Tagger*	○ 0	64.6	47.6 (-17.0)	42.7 (-21.9)	-
LayoutLM-Base Tagger ^b	○ 500K	69.9	-	-	-
LayoutLM-Base Tagger*	○ 11M	78.9	72.5 (-6.4)	70.2 (-8.7)	-
SPADE \diamond [†]	× 0	71.6	70.5 (-1.1)	72.0 (+0.4) [§]	41.3

^aJaume et al. (2019). ^bFrom Xu et al. (2019).

* The source code from <https://github.com/microsoft/unilm/tree/master/layoutlm>.

[§] The separation of long input text (> 512) into multiple independent inputs introduces small difference in F_1 .

[†]Five encoder layers are used for computational efficiency.

Ablation study We probe the role of each component of SPADE \diamond via ablation study (Table 5). The performance drops dramatically upon the removal of the relative coordinate information of tokens in the self-attention layer, highlighting its importance in the serializer-free encoder (2nd row). When the absolute coordinates are used in the input instead of the relative coordinates, F_1 drops by 6.9% (3rd row). Finally, 2.6% drop in F_1 is observed upon the removal of the data augmentation during training (4th row).

Table 5: Ablation study on CORD dataset.

Model	F_1
SPADE \diamond [†]	84.5
(-) relative coordinate	10.5 (-74.0)
(-) relative coordinate (+) absolute coordinate	78.6 (-6.9)
(-) data augmentation	81.9 (-2.6)

[†] Five encoder layers are used for computational efficiency.

7 Conclusion

We present SPADE \diamond , a spatial dependency parser that can extract highly structured information from documents that have complex layouts. By formulating document IE as a spatial dependency graph construction problem, we provide a powerful unified framework that can extract hierarchical information without feature engineering. We empirically demonstrate the effectiveness of our model over various real-world documents—receipts, name cards, and invoices—and in a popular form understanding task.

Acknowledgments

We thank Geewook Kim for the critical comments on the manuscript and Teakgyu Hong and Sungrae Park for the helpful discussions on FUNSD experiments.

References

- Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoon Yun, Seong Joon Oh, and Hwalsuk Lee. 2019a. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *ICCV*.
- Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoon Yun, and Hwalsuk Lee. 2019b. Character region awareness for text detection. In *ICCV*, pages 9365–9374.
- Jason P.C. Chiu and Eric Nichols. 2016. **Named entity recognition with bidirectional LSTM-CNNs**. *TACL*, 4:357–370.
- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *ACL*.
- Timo I. Denk and Christian Reisswig. 2019. **BERT-grid: Contextualized Embedding for 2D Document Representation and Understanding**. *arXiv e-prints*, page arXiv:1909.04948.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *NAACL*.
- Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *ACL*.
- Lukasz Garncarek, Rafal Powalski, Tomasz Stanislawek, Bartosz Topolski, Piotr Halama, and Filip Gralinski. 2020. **LAMBERT: layout-aware language modeling using BERT for information extraction**. *arXiv e-prints*.
- Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2021. **BROS: A pre-trained language model for understanding texts in document**.
- Wonseok Hwang, Seonghyeon Kim, Jinyeong Yim, Minjoon Seo, Seunghyun Park, Sungrae Park, Junyeop Lee, Bado Lee, and Hwalsuk Lee. 2019. Post-ocr parsing: building simple and robust parser via bio tagging. In *Workshop on Document Intelligence at NeurIPS 2019*.
- Wonseok Hwang, Hyunji Lee, Jinyeong Yim, Geewook Kim, and Minjoon Seo. 2021. **Cost-effective end-to-end information extraction for semi-structured document images**. *arXiv e-prints*, abs/2104.08041.
- Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *ICDAR-OST*.
- Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2D documents. In *EMNLP*.

- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL*.
- Bill Yuchen Lin, Ying Sheng, Nguyen Vo, and Sandeep Tata. 2020. Freedom: A transferable neural architecture for structured information extraction on web documents. In *KDD*.
- Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph convolution for multimodal information extraction from visually rich documents. In *NAACL*.
- Colin Lockard, Prashant Shiralkar, and Xin Luna Dong. 2019. OpenCeres: When open information extraction meets the semi-structured web. In *ACL*.
- Colin Lockard, Prashant Shiralkar, Xin Luna Dong, and Hannaneh Hajishirzi. 2020. ZeroShotCeres: Zero-shot relation extraction from semi-structured webpages. In *ACL*.
- Xuezhe Ma and Eduard Hovy. 2020. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *ACL*.
- Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James B. Wendt, Qi Zhao, and Marc Najork. 2020. Representation learning for information extraction from form-like documents. *ACL*.
- Rasmus Berg Palm, Ole Winther, and Florian Laws. 2017. Cloudscan - A configuration-free invoice analysis system using recurrent neural networks. *CoRR*.
- Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: A consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.
- Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *TACL*, 5:101–115.
- Rafal Powalski, Lukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michal Pietruszka, and Gabriela Palka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. *CoRR*, abs/2102.09550.
- Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay. 2019. GraphIE: A graph-based framework for information extraction. In *NAACL*.
- Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.
- Johannes Rausch, Octavio Martinez, Fabian Bissig, Ce Zhang, and Stefan Feuerriegel. 2019. Docparser: Hierarchical structure parsing of document renderings. *CoRR*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NeurIPS*.
- Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. 2021. Towards robust visual information extraction in real world: New dataset and novel solution. In *AAAI*.
- Mengxi Wei, Yifan He, and Qiong Zhang. 2020. Robust Layout-aware IE for Visually Rich Documents with Pre-trained Language Models. In *SIGIR*.
- Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2020. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. *arXiv e-prints*.
- Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2019. Layoutlm: Pre-training of text and layout for document image understanding. In *KDD*.
- Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. 2020. PICK: Processing Key Information Extraction from Documents using Improved Graph Learning-Convolutional Networks. In *ICPR*.
- Luke S Zettlemoyer and Michael Collins. 2012. Learning to map sentences to logical form: Structured classification with probabilistic categorical grammars.
- Xiaohui Zhao, Zhuo Wu, and Xiaoguang Wang. 2019. CUTIE: learning to understand documents with convolutional universal text information extractor. *arXiv e-prints*.
- Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno-Yepes. 2019. Image-based table recognition: data, model, and evaluation. *arXiv e-prints*.

A Appendices

A.1 Dataset

A.1.1 Dataset collection

The internal datasets RECEIPT-IDN, NAMECARD and INVOICE are annotated by the crowd through an in-house web application following (Park et al., 2019; Hwang et al., 2019). First, each text segment is labeled (bounding box and the characters inside) for the OCR task. The text segments are further grouped according to their field types by the crowds. For RECEIPT-IDN and INVOICE, additional group-ids are annotated to each field for inter-grouping of them. The text segments placed on the same line are also annotated through row-ids. For quality assurance, the labeled documents are cross-inspected by the crowds.

A.1.2 CORD, CORD+, CORD++, and CORD-M for receipt IE

CORD and their variant consist of 30 information categories such as `menu_name`, `count`, `unit_price`, `price`, and `total_price` (Table 6). The fields are further grouped and forms the information layer at a higher level.

A.1.3 RECEIPT-IDN for receipt IE

RECEIPT-IDN is similar to CORD but includes more diverse information categories (50) such as `store_name`, `store_address`, and `payment_time` (Table 6).

A.1.4 NAMECARD for name card IE

NAMECARD consists of 12 field types, including `name`, `company_name`, `position`, and `address` (Table 6). The task requires grouping and ordering of tokens for each field. Although there is only a single information layer (field), the careful handling of complex spatial relations is required due to the large degree of freedom in the layout.

A.1.5 INVOICE for invoice IE

INVOICE consists of 62 information categories such as `item_name`, `count`, `price_with_tax`, `item_price_without_tax`, `total_price`, `invoice_number`, `invoice_date`, `vendor_name`, and `vendor_address` (Table 6). Similar to receipts, their hierarchical information is represented via inter-field grouping.

A.1.6 FUNSD for general form understanding

FUNSD form understanding task consists of two sub tasks: entity labeling (ELB) and entity linking (ELK). In ELB, tokens are classified into one of four fields—header, question, answer, and other—while doing serialization of tokens within each field. Both subtasks assume that the input tokens are perfectly serialized with no OCR error. To emphasize the importance of correct serialization in the real-world, we prepare two variant of ELB tasks: ELB-R and ELB-S. In ELB-R, the whole documents are randomly rotated by a degree of -20° – 20° and the input tokens are serialized using rotated y-coordinates. In ELB-S task, the input tokens are randomly shuffled. In both tasks, the relative order of the input tokens within each field remain unchanged. In ELK task, tokens are linked based on their key-value relations (inter-grouping between fields). For example, each “header” is linked to the corresponding “question”, and “question” is paired with the corresponding “answer”.

Table 6: The representative fields of the datasets.

Dataset	representative fields and their numbers
CORD,CORD+, CORD++,CORD-M	<code>menu_name</code> (2572), <code>count</code> (2357), <code>unit_price</code> (737), <code>price</code> (2559), <code>total_price</code> (974)
RECEIPT-IDN	<code>menu_name</code> (28832), <code>menu_count</code> (27132), <code>menu_unitprice</code> (11530), <code>menu_price</code> (28028), <code>total_price</code> (10284), <code>store_name</code> (9413), <code>payment_time</code> (9817)
NAMECARD	<code>name</code> (25917), <code>company_name</code> (24386), <code>position</code> (22848), <code>address</code> (26018)
INVOICE	<code>item_name</code> (2761), <code>count</code> (1950), <code>price_with_tax</code> (781), <code>price_without_tax</code> (2230), <code>total_price</code> (844), <code>invoice_number</code> (803), <code>invoice_date</code> (987), <code>vendor_name</code> (993), <code>vendor_address</code> (993).
FUNSD ^a	<code>header</code> (563), <code>question</code> (4343), <code>answer</code> (3623), <code>other</code> (1214)

^a From (Jaume et al., 2019).

A.2 Evaluation metric

During calculation of F_1 for parses, the difference between prediction and ground truth is not counted in `store_name`, `menu_name`, and `item_name` fields in receipt and invoice when the edit distance (ED) is less than 2 or when the $ED/gt\text{-string-length} \leq 0.4$. Also, in Japanese documents, white spaces are ignored.

In the FUNSD form understanding task, we measure entity labeling (ELB) and entity linking (ELK) scores following (Jaume et al., 2019). ELB measures the field classification accuracy of already “perfectly” serialized tokens of each field (words group), whereas ELK measures the inter-grouping accuracy between word groups. As SPADE Δ does both the serialization of the fields and grouping between fields simultaneously, we do not feed the serialized tokens into SPADE Δ but only use the oracle information to indicate the first text node of each field from the predicted graph. These text nodes effectively represent entire fields and are used for the evaluation.

A.3 The score for the dependency relation prediction

Table 7: The score for the dependency relation prediction. s and g stand for `rel-s` and `rel-g`.

Model	rel	Precision					Recall					F_1				
		CO	RI	NC	INV	FU	CO	RI	NC	INV	FU	CO	RI	NC	INV	FU
Δ - TCA	s	96.4	97.7	90.7	97.4	60.6 \dagger	97.1	98.8	92.0	98.3	63.7\dagger	96.8	98.3	91.3	97.8	62.2 \dagger
Δ - TCA	g	87.8	91.1	-	86.7	41.1 \dagger	90.1	93.8	-	88.0	34.4\dagger	88.9	92.4	-	87.3	37.4 \dagger
Δ	s	96.8	97.8	91.9	97.6	70.4\dagger	97.1	98.8	91.3	98.2	59.8 \dagger	96.9	98.3	91.6	97.9	64.6\dagger
Δ	g	89.9	92.2	-	88.6	49.7\dagger	89.2	93.1	-	86.3	30.5 \dagger	89.6	92.7	-	87.4	37.8\dagger
UB-NO-SER	s	100	100	100	100	-	32.7	31.3	57.7	18.8	-	49.3	47.7	73.1	31.7	-
UB-NO-SER	g	0	0	0	0	-	0	0	0	0	-	0	0	0	0	-

\dagger Five encoder layers are used instead of twelve for computational efficiency.

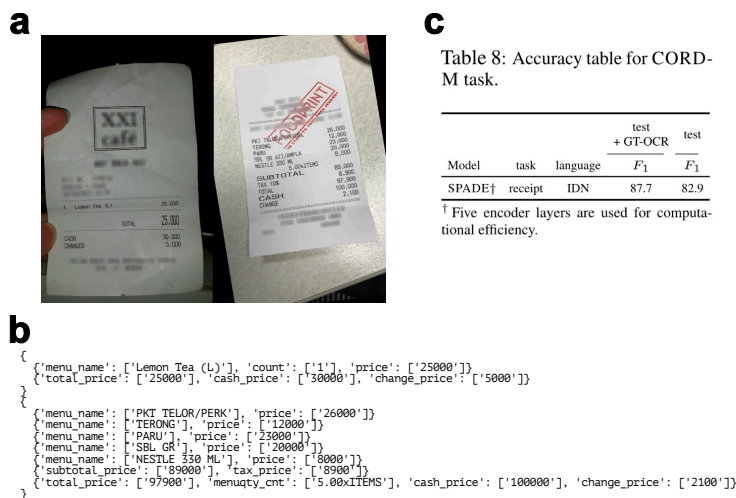


Figure 4: The example of a receipt image from CORD-M (a), the predicted parse (b), and the accuracy table (c).

