Utility_Bill Labeling Instructions

September 2022

Confidential + Proprietary

Labelling Guidance

Getting Started

- Start by annotating a small number of documents
- Review them, correct any mistakes, update the labeling guidance to make sure data is labeled consistently.

Choosing the right data/model

- A general rule is to uptrain with customer production traffic. The closer the training distribution is to production traffic, the better the performance.
- Depending on your customer's requirements, consider utilizing more than one processor in order to support multiple languages
- One model for every region (group of untrained languages) vs a universal model

Dataset Size

- Training - at least 1000 per language
- Test - at least 200 per language

Confidential + Proprietary

Multi-Occurrence entities

There are 2 occurrence types that we are relevant to utility_bill labeling:

- **`OPTIONAL_ONCE`**: This entity is expected to have no more than **one** ground truth value.
- **`OPTIONAL_MULTIPLE`**: This entity can have any number of ground truth values.

For utility_bill, **line item,service and vat** entities are **`OPTIONAL_MULTIPLE`**, while all other entities are **`OPTIONAL_ONCE`**.

Note that while **`OPTIONAL_ONCE`** entities can still appear multiple times with the same value. In this case, we recommend to label all appearances, except for “easy” entities (in the current entity list, only **`currency`**).

Confidential + Proprietary

Optical Character Recognition (OCR) Errors

Drawing bounding boxes around entities will automatically populate a text captured through OCR. In some cases, the captured text is not accurate and doesn't reflect what is on the page, especially for documents that are blurry or of low resolution. These values should be corrected manually through "Value" input text.

If a value doesn't exist in the document, there is no need to label it.

Confidential + Proprietary

All Entities

- unit_of_measure
- service_type
- usage
- amount_due
- currency
- due_date
- invoice_date
- invoice_id
- late_fee_amount
- line_item/amount
- line_item/unit_price
- line_item/quantity
- line_item/service_end_date
- line_item/service_start_date
- line_item/description
- line_item/unit
- line_item/usage
- prior_amount_due
- prior_paid_amount
- receiver_name
- receiver_address
- receiver_phone
- receiver_email
- service_address
- service_end_date
- service_start_date
- service_id
- supplier_account_number
- supplier_address
- supplier_name
- supplier_phone
- supplier_email
- supplier_website
- total_amount
- total_tax_amount
- line_item/unit_of_measure
- line_item/product_code
- line_item/tax_amount
- net_amount
- adjusted_amount
- remit_to_name
- remit_to_address
- supplier_registration
- supplier_tax_id
- vat/amount
- vat/tax_amount
- vat/tax_rate
- receiver_tax_id
- delivery_date

Confidential + Proprietary

Utility Bill Labeling Instruction

Confidential + Proprietary

Invoice entities

Invoice_date:
This field refers to the invoice generation date. Text in the invoice that refers to/qualifies the invoice date. Example: “Dated:”, “Invoice Date”
Special Instructions:

1. In some cases, only month & year may be mentioned, instead of the full date. Example “Dec 2019”. Please consider such values as valid and annotate them.
2. Just include the actual date text in the value bounding box. Don’t include words like “Date” or “Date of Invoice”

Invoice_id:
Text in the invoice that refers to/qualifies the Invoice Identifier. Example: “Invoice Id”, “Invoice #”
The actual invoice ID. Include any prefix/suffix associated with the invoice id. Example- “INV-23”

Confidential + Proprietary

line_item entities

line_item/amount :
One header of the items table referring to the billed amount for a particular line item. Examples: “Amount”, “Total” the actual amount billed for a particular item in the items table. Usually, one value per line item in the items table. Examples: “20$”, “65.4 GBP”

line_item/unit_price:

Usually it’s one header of the items table describing the price per unit of item being ordered. Examples: “Price / hr”, “Hourly Price”it is the actual price per unit of the item being ordered. There are usually many values in a single image, one corresponding to each item. Examples: “70”, “50 $/hr”, “20 $/pound”

line_item/quantity:

Usually it’s one header of a item table referring to quantity. Examples: “Number of units”, “Qty”, “Quantity”. it is the actual number of units ordered. There are usually many values in a single image, one corresponding to each item. Examples: “10”, “20 bags”, “30 lbs”

Confidential + Proprietary

line_item entities

line_item/description:

usually a table header referring to descriptions or names of the item. Examples: , “Product”, “Description”, “Item Description”. line_item/description is a actual description of the item. There are usually many values in a single image.Examples: “Laptop”, “Paper bundle”, “HP Headphones”

line_item/service_start_date:

Start date for the utility services charged in the utility statement.

line_item/service_end_date:

End date for the utility services charged in the utility statement.

line_item/unit:

Usually it’s one header of the items table describing a column referring to the measurement units of the item. Examples: “Unit”, “Hrs/Qty” it is the measurement unit of the item being ordered. There are usually many values in a single image, one corresponding to each item. Examples: “lb”, “pounds”, “kg”, “hours”

Confidential + Proprietary

line_item entities

line_item/unit_of_measure:
It is the measurement unit of the item being ordered.

line_item/product_code:

Usually it's one header of the items table referring to a product identifier. Examples: "ID", "Product ID", "Serial Number", "FSN #", "Item Code". It is the actual product ID/code. There are usually many values in a single image, one corresponding to each item. Examples: "BN-150", "75", "B-108".

line_item/tax_amount:

Tax amount charged for a particular utility line item (or standalone/separate tax line item included).

line_item/usage:

Usage amount shown in the utility line item. It's usually associated with an unit of measure shown in the statement.

Confidential + Proprietary

receiver entities

receiver_address:

Text in the invoice that refers to/qualifies the receiver's address. full address of the organization/person/entity to which the invoice is made or billed. Do not include the fax/phone number/email in this field. Example: "1345 Main Street, Chicago IL 60007".

receiver_phone:

Phone number of the receiver (customer).

receiver_name:

Text in the invoice that refers to/qualifies the name of the receiver organization/person. name of the organization/person/entity to which the invoice is made or billed. Examples: "Google Ireland Limited". Do not include the address/phone number in this field.

receiver_email: Email id of the receiver.

Confidential + Proprietary

service entities

service_id:

One key (found as a header in a table or as standalone) referring to the specific measurement device, e.g., meter, circuit. the actual identifier for the device

service_start_date:
Start date for the utility services charged in the utility statement.

service_end_date:
End date for the utility services charged in the utility statement.

service_address:

service address for a particular line item. It’s usually related to a particular housing unit / apartment

Confidential + Proprietary

Supplier entities

supplier_account_number:

Supplier's account number for the customer related to the utility line item.

supplier_name:

Text in the invoice that refers to/qualifies the name of the organization/person, that is preparing the invoice. Example: "From", Name of the organization/person/entity which has generated the invoice/supplied the items. Examples: "Deems Uk Ltd", "Fab India Pvt Ltd". Do not include the address/phone number in this field.

supplier_address:

Text in the invoice that refers to/qualifies the supplier's address. full address of the organization/person/entity which has issued the invoice/supplied the items. Do not include the fax/phone number/email in this field. Example: "1345 Main Street, Chicago IL 60007"

supplier_phone:

Phone number of the supplier (utility company).

supplier_email:

email address of the supplier

Confidential + Proprietary

Supplier entities

supplier_registration:
Business registration of the supplier (utility company).

supplier_tax_id:

Text in the invoice that refers to/qualifies the tax id of the supplier. the actual tax id of the supplier.

supplier_website:

website of the supplier.

Confidential + Proprietary

Miscellaneous entities

total_amount:
Text in the invoice that refers to/qualifies the total billed amount in the invoice. Example - “Total”, “Total Amount”

total_tax_amount :
Text in the invoice that refers to/qualifies the total tax amount mentioned in the invoice. Eg. “VAT”, “Sales Tax”, “Excise”, “Custom duty” the monetary tax amount mentioned in the invoice. Do NOT include currency symbols or code in this entity.

Confidential + Proprietary

Miscellaneous entities

net_amount:

The actual text indicating the monetary amount. Do NOT include currency symbols or code in this entity.

adjusted_amount:
Amount shown on the utility statement following prior adjustments.

remit_to_name:
Text identifying the entity, to whom payment must be remitted to. Examples: “Remit To”, “Send Payment to Address”. full name of entity to whom the payment should be remitted to. Examples: “Google LLC”.

prior_paid_amount:

Amount paid by the customer since the last utility statement.

prior_amount_due: Amount which is due previously

Confidential + Proprietary

Miscellaneous entities

vat/amount:

It is the actual amount. Examples: "6,49". Associated key (column header) example: "Base Taxable","Amount".

vat/tax_rate:

It is the actual tax rate. Examples: "13.00". Associated key (column header) example: "%TVA".

vat/tax_amount:

It is the actual tax amount. Examples: "6,49". Associated key (column header) example: "TVA", "Mountant TVA", "VAT Amount" (Notice this column is not for vat/amount).

receiver_tax_id:

Tax id of the receiver.

delivery_date:

Text in the invoice that refers to/qualifies the date when the items will be/were delivered. Example: “Delivery date”.the actual delivery date mentioned in the invoice.

Confidential + Proprietary

Miscellaneous entities

remit_to_address: Text referring to the payment remit address for the invoice. Examples: “Remit To”, “Send Payment to Address”. full address where the payment should be sent to. Do not include the fax/phone number/email in this field. Examples: “1345 Main Street, Chicago IL 60007”. Do not include bank related information.

due_date:

Text in the invoice that refers to/qualifies the due date of the invoice. Example: “Due date”, “Due by”, “Pay by”. the actual date when the customer should settle the invoice payment. Example: “12/31/2019”, “Dec 2019”, “5th May 2020”

late_fee_amount: Enter the late fee amount

Confidential + Proprietary

Miscellaneous entities

unit_of_measure:

Units in which service is measured.

usage: Service usage.

amount_due:

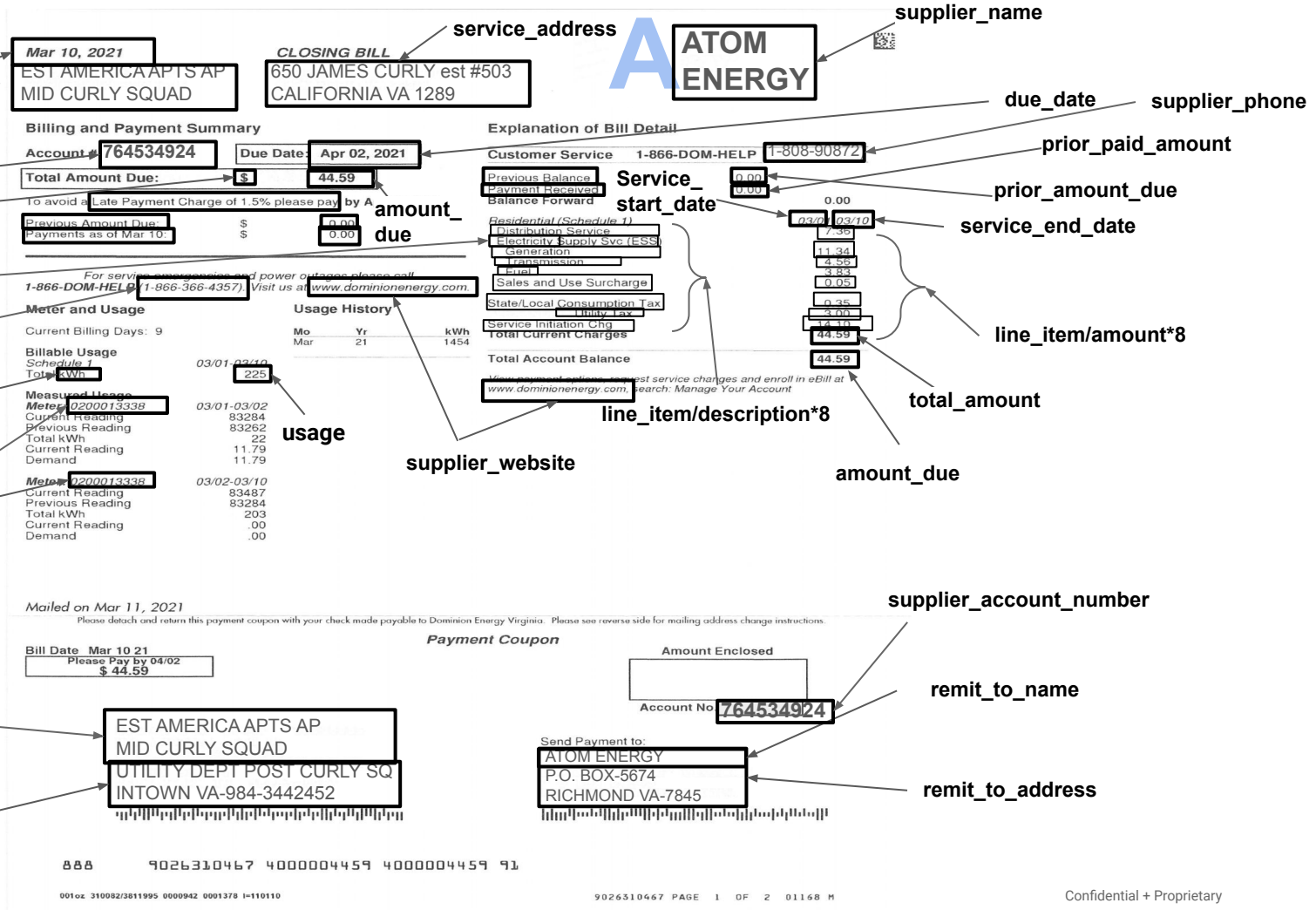
Text in the invoice that refers to/qualifies the total billed amount in the invoice. Example - “Amount due”, “Total due”, "Please Pay" and "Pay this amount".The actual text indicating the monetary amount. Do NOT include currency symbols or code in this entity.

currency:

It is the actual currency for the amount. Examples: “USD”, “GBP”, “$”, “€”, “US $”, it is text referring to the currency type eg. “Currency”, “FX”.

Confidential + Proprietary

Example 1



Confidential + Proprietary

Feedback

1. Supplier_phone is not labeled(two places) and is not a legitimate phone number(top right corner), please correct it.

Confidential + Proprietary