Expense Labeling Instructions

chenblair@google.com

September 2022
General Labeling Instructions
Labeling 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
Multiple Occurrence Handling

There are 2 occurrence types that we are relevant to expense 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 expenses, **line item** 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**).
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.
Expense Entity Definitions
Expense Parser Entities

In total, there are **22** entity types to be labeled. These types can be broken down broadly into **5** categories (+ miscellaneous).

<table>
<thead>
<tr>
<th><strong>Supplier</strong></th>
<th><strong>Timestamp</strong></th>
<th><strong>Line Item</strong></th>
<th><strong>Amount</strong></th>
<th><strong>Payment</strong></th>
<th><strong>Miscellaneous</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>supplier_name</td>
<td>purchase_time</td>
<td>line_item/amount</td>
<td>total_amount</td>
<td>currency</td>
<td>traveler_name</td>
</tr>
<tr>
<td>supplier_address</td>
<td>receipt_date</td>
<td>line_item/description</td>
<td>total_tax_amount</td>
<td>payment_type</td>
<td>reservation_id</td>
</tr>
<tr>
<td>supplier_phone</td>
<td>start_date</td>
<td>line_item/quantity</td>
<td>net_amount</td>
<td>credit_card_last_four_digits</td>
<td></td>
</tr>
<tr>
<td></td>
<td>end_date</td>
<td>line_item/product_code</td>
<td>tip_amount</td>
<td>payment_authorization_id</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>line_item/transaction_date</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplier Entities

A **supplier** is the business/individual who provided the goods or services involved in the expense.

**supplier_name** - Name of the supplier

**supplier_address** - Address of the supplier. In a physical transactions, this will often be the location at which the transaction was conducted.

**supplier_phone** - Phone number of the supplier
Timestamp entities identify when a transaction occurred. Some entities only apply to continuous transactions that occur over the span of days such as car rental or hotel stays.

**purchase_time** - Clock time of transaction

**receipt_date** - Date of transaction

**start_date** - Start date of a continuous transaction

**end_date** - End date of a continuous transaction
Line Item Entities

**Line item** entities appear in the itemized list of what was purchased. There will probably be multiple of each line_item entity per receipt.

- **line_item/amount** - Total cost of line item
- **line_item/description** - Description of line item
- **line_item/quantity** - Quantity of line item purchased
- **line_item/product_code** - Unique identifying code for line item
- **line_item/transaction_date** - Date of line item purchase
Amount entities identify aggregate payment amounts owed in the transaction

**total_amount** - Total payment owed

**total_tax_amount** - Total tax owed

**net_amount** - Total payment, not including taxes, tips, or additional fees

**tip_amount** - Total tip paid
Payment entities detail the transaction payment and how it was conducted.

**currency** - Currency of the transaction

**payment_type** - What form of payment was used (VISA, AMEX, Cash, etc.)

**payment_authorization_id** - ID that uniquely identifies the transaction

**credit_card_last_four_digits** - If a credit/debit card was used, the last four digits of the card number


**traveler_name** - Name of the traveler on an expense involving transportation of lodging

**reservation_id** - Confirmation number given to customer at time of booking

Thanks for renting from Car Renters! We hope to see you again soon.