Probabilistic Programming for Voucher Information Extraction
Preliminary Practical Experiences
Al-Sibahi, Ahmad Salim; Hamelryck, Thomas Wim; Henglein, Fritz

Publication date:
2018

Document version
Early version, also known as pre-print

Citation for published version (APA):
Probabilistic Programming for Voucher Information Extraction

Ahmad Salim Al-Sibahi
University of Copenhagen/Skanned.com

Thomas W. Hamelryck
University of Copenhagen

Fritz Henglein
University of Copenhagen

Introduction to Skanned.com

Skanned.com provides a Voucher Scanning service for extracting information from vouchers like product lines, total amounts, payment date, sender and recipient.

Vouchers vary heavily in size, layout, purpose and content; the scan quality is occasionally suboptimal. Probabilistic programming provides an opportunity to:

- Combine domain knowledge and machine learning to effectively extract features in a systematic fashion.
- Quantify confidence in results, which is important for manual validation.

Skanned.com's Pipeline

- OCR Optical Character Recognition extracts textboxes from PDFs.
- Feature Extractors extract information from the text boxes.

Finding Features w/Keywords

Features are usually located around identifying keywords. Keywords can be positive or negative depending on the feature to be found.

<table>
<thead>
<tr>
<th>Total Amount Excl. VAT</th>
<th>23613.00 DKK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total VAT</td>
<td>5903.25 DKK</td>
</tr>
<tr>
<td>Total Amount</td>
<td>29516.25 DKK</td>
</tr>
</tbody>
</table>

Probabilistic model below tries to infer a latent score $r$ from the vector of observed angles $\theta^+$ and distances $d^+$ from positive keywords to potential target features.

\[
\begin{align*}
    r & \sim \mathcal{B}(0.5,0.5) \\
    \mu_1^+ & = (0, \frac{\pi}{2}) \\
    w_1^+ & = (0.7, 0.3) \\
    \mu_2^+ & = (-\frac{\pi}{2}, \frac{\pi}{4}, \frac{3\pi}{4}) \\
    w_2^+ & = (0.5, 0.2, 0.3) \\
    \bar{\theta}^+ | r & \sim \sum_{j=1}^{2} \alpha_j \gamma \sum_{i=1}^{w_j} \nu_j \gamma \mathcal{M}(\mu_j^+, \frac{4}{\pi}) \\
    \bar{d}^+ | r & \sim \tilde{\gamma}_1 \gamma \mathcal{N}(500) + \tilde{\gamma}_2 \gamma \mathcal{N}(1500,1000)
\end{align*}
\]

Practical Experiences

Sampling

- Ease of use ✓ Precision ❌ Scalability

Variational Inference

- Scalability ✓ Set-up ✓ Precision

GPU Support

- Discrete Latents ✦ Ease of use ❌ Precision

The first author would like to thank Dan Rose and the Scanner Team at Skanned.com (Bjørn Kaae, Toke Reines, Danni Dromi) for discussions and feedback on the work!