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):
Al-Sibahi, A. S., Hamelryck, T. W., & Henglein, F. (2018). *Probabilistic Programming for Voucher Information Extraction: Preliminary Practical Experiences*. Poster session presented at PROBPROG 2018 - The International Conference on Probabilistic Programming, Cambridge, MA, United States.

Probabilistic Programming for Voucher Information Extraction

Preliminary Practical Experiences

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

Thomas W. Hamelryck
University of Copenhagen

Fritz Henglein
University of Copenhagen

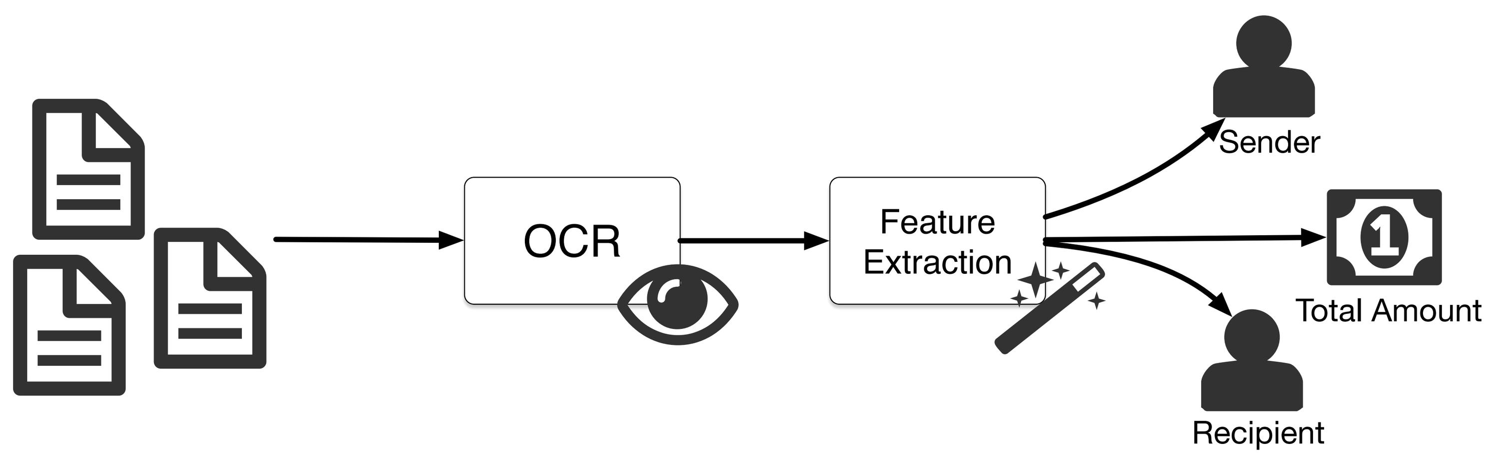
Introduction to Skanned.com

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

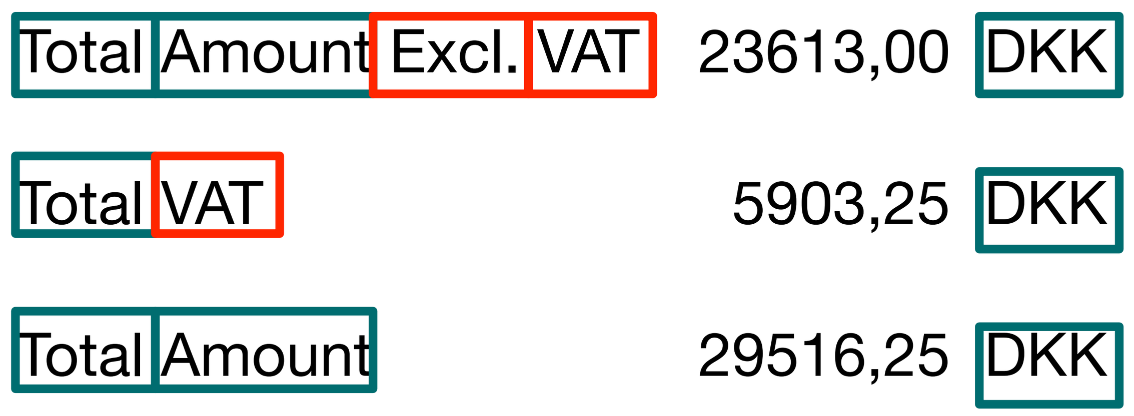


- **OCR** Optical Character Recognition extracts textboxes from PDFs.
- **Feature Exctractors** extract information from the text boxes.

skanned
University of Copenhagen
Faculty of Science

Finding Features w/Keywords

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



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

$$r \sim \mathcal{B}(0.5, 0.5) \qquad \tilde{r} = (r, 1-r)$$

$$w_1^{+} = (0.7, 0.3) \qquad \mu_1^{+} = (0, \frac{\pi}{2})$$

$$w_2^{+} = (0.5, 0.2, 0.3) \qquad \mu_2^{+} = (-\frac{\pi}{2}, \frac{\pi}{4}, \frac{3\pi}{4})$$

$$\overrightarrow{\theta^{+}} \mid r \overset{\text{iid}}{\sim} \sum_{j=1}^{2} \tilde{r}_j \sum_{i=1}^{|\mu_j^{+}|} w_{j,i}^{+} \text{vM}(\mu_{j,i}^{+}, \frac{4}{\pi})$$

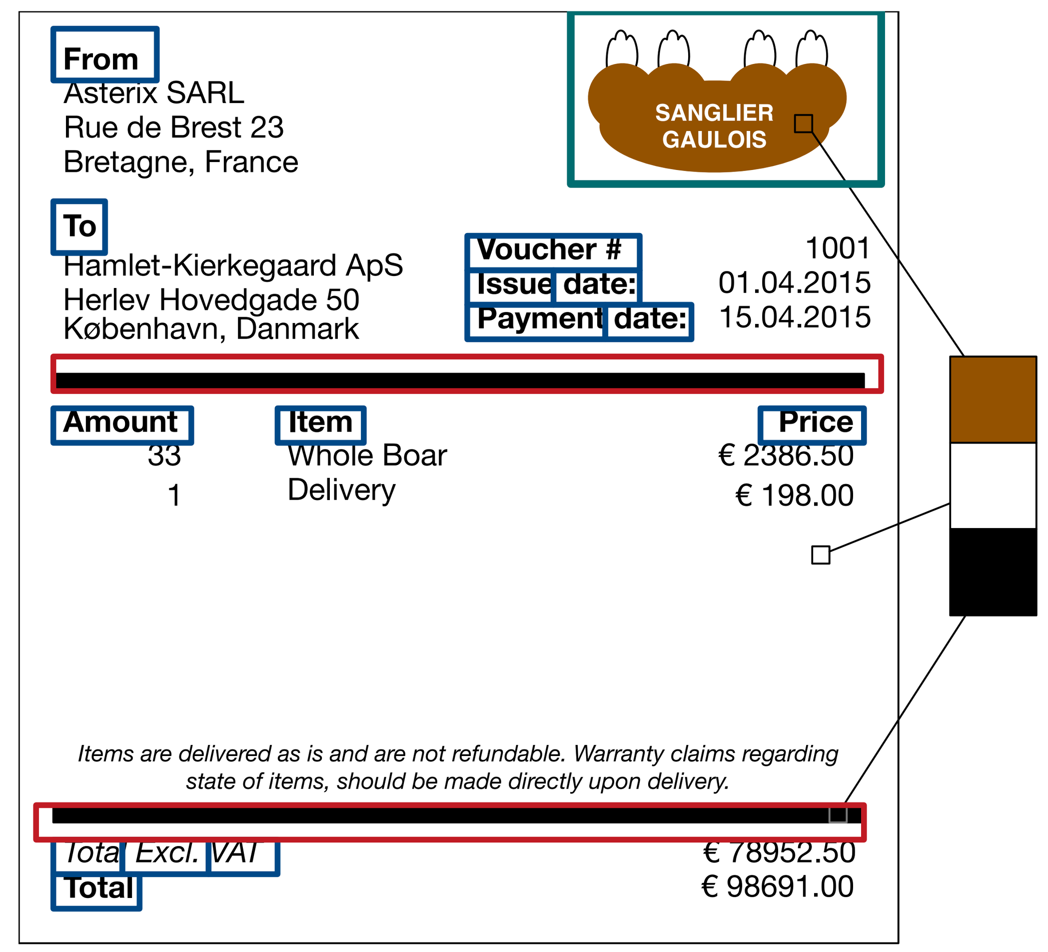
$$\overrightarrow{d^{+}} \mid r \overset{\text{iid}}{\sim} \tilde{r}_1 \mathcal{N}^{+}(500) + \tilde{r}_2 \mathcal{N}(1500, 1000)$$

Evaluating extended version on 1000 vouchers:

- **80%** of the time the expected score found the target feature
- **99%** of the time it was within confidence interval

Voucher Grouping

To provide more accurate models, to partition the voucher into groups of similar layout and style. We rely on probabilistic Latent Dirichlet Allocation (LDA) to perform the grouping, using visual (colors, lines) and textual cues (keywords).



For 1000 sample vouchers we achieved 21 topics, and our next goal is to rely on these topics to construct more precise feature extraction models.

Practical Experiences

Sampling

✓ Ease of use ✓ Precision ✗ Scalability

Variational Inference

✓ Scalability ✦ Set-up ✦ Precision

GPU Support

✦ Ease of use

Discrete Latents

✗ 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!