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.

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

\[
\begin{align*}
\text{Total Amount Excl. VAT} & \quad 23613.00 \text{ DKK} \\
\text{Total VAT} & \quad 5903.25 \text{ DKK} \\
\text{Total Amount} & \quad 29516.25 \text{ DKK}
\end{align*}
\]

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.

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

- **Easy 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!