SPLIT-CORRECTNESS IN INFORMATION EXTRACTION

Johannes Doleschal - Benny Kimelfeld – Wim Martens
Yoav Nahshon - Frank Neven
Motivation

**Information Extraction**

**Input:** Huge text documents, e.g. from server logs or financial reports

**Task:** Convert the data into a structured relational format
Information Extraction

Naive Extraction
Information Extraction

Naive Extraction

Parallel Extraction

Possible splits $\mathcal{K}$:
- Text documents into paragraphs or sentences,
- error logs into exceptions,
- server logs into HTTP messages, ...
Information Extraction

Naive Extraction

\[
\begin{array}{c}
\text{Naive Extraction} \\
\text{\rightarrow} \\
\end{array}
\]

Parallel Extraction

\[
\begin{array}{c}
\text{Parallel Extraction} \\
\text{\rightarrow} \\
\end{array}
\]

\[
\begin{array}{c}
\text{\text{\ldots}} \\
\text{\rightarrow} \\
\text{\rightarrow} \\
\text{n} \\
\text{n} \\
\end{array}
\]
Information Extraction

Naive Extraction

Parallel Extraction

Split-Correctness in Information Extraction – PODS'19
More Formally

Information Extractor $\rightarrow$

$\rightarrow(\mathcal{E}) = \mathbb{P}$

Splitter $\mathcal{X}$

$\mathcal{X}(\mathcal{E}) = \mathcal{E}_1, \ldots, \mathcal{E}_n$

Evaluating Extractor $\rightarrow$ on Splitter $\mathcal{X}$

$\rightarrow(\mathcal{X}(\mathcal{E})) = \bigcup_{1 \leq i \leq n} \rightarrow(\mathcal{E}_i) = \mathbb{P}$
Central Question

Naive Extraction → \textit{global}(\mathcal{E})

Parallel Extraction → \textit{local}(\mathcal{E}(\mathcal{S}))

Extractor → \textit{global}

Splitter \mathcal{S}

Extractor → \textit{local}

Basically just a Union Operator
Main Problems

Split-Correctness
Input: Information Extractors $\rightarrow_{\text{global}}$, $\rightarrow_{\text{local}}$, Splitter $\kappa$
Question: Is $\rightarrow_{\text{global}}(\mathcal{D}) = \rightarrow_{\text{local}}(\kappa(\mathcal{D}))$, for all documents $\mathcal{D}$?

Self-Splittability
Input: Information Extractor $\rightarrow$, Splitter $\kappa$
Question: Is $\rightarrow(\mathcal{D}) = \rightarrow(\kappa(\mathcal{D}))$, for all documents $\mathcal{D}$?
Main Problems

Input: Information Extractor $\rightarrow_{\text{global}}$, Splitter $\kappa$

Question: Is there an Information Extractor $\rightarrow_{\text{local}}$, such that $\rightarrow_{\text{global}}(\mathcal{G}) = \rightarrow_{\text{local}}(\kappa(\mathcal{G}))$, for all documents $\mathcal{G}$?

Splittability

Parallel Extraction $\rightarrow_{\text{local}}(\kappa(\mathcal{G}))$

Naive Extraction $\rightarrow_{\text{global}}(\mathcal{G})$
Regular Document Spanners/Splitters

Information Extractors: Regular document spanners
Finite state automata with variable operations. Essentially, regular expressions with capture variables, closed under relational algebra ($\cup, \pi, \bowtie$).

Splitters: Regular document Splitters
Unary regular document spanners

E.g. sentence extractor
A splitter is disjoint, if every part of a document is in at most one output document.

**Disjoint Splitter**
- e.g. Sentence segmentation
  - N-Grams
  - Split into paragraphs

**Non-Disjoint Splitter**
- e.g. Pairs of consecutive sentences
Main Results

**Self-Splittability**

- **Input:** Information Extractor $\rightarrow$, Splitter $\mathfrak{K}$
- **Question:** Is $\operatorname{Is} \rightarrow(\mathcal{I}) = \rightarrow(\mathfrak{K}(\mathcal{I}))$, for all documents $\mathcal{I}$?

**Split-Correctness**

- **Input:** Information Extractors $\rightarrow_{\text{global}}$, $\rightarrow_{\text{local}}$, Splitter $\mathfrak{K}$
- **Question:** Is $\operatorname{Is} \rightarrow_{\text{global}}(\mathcal{I}) = \rightarrow_{\text{local}}(\mathfrak{K}(\mathcal{I}))$, for all documents $\mathcal{I}$?

---

**Theorem 5.1 and Theorem 5.16**

Self-Splittability and Split-Correctness for regular document spanners are PSPACE-complete

**Theorem 5.7 and Theorem 5.17**

Self-Splittability and Split-Correctness for disjoint splitters and regular document spanners in normal form* is in polynomial time

* deterministic + functional

Split-Correctness in Information Extraction – PODS'19
Main Results

**Splittability**

Input: Information Extractor $\rightarrow_{\text{global}}$, Splitter $\mathcal{S}$
Question: Is there an Information Extractor $\rightarrow_{\text{local}}$, such that $\rightarrow_{\text{global}}(\mathcal{A}) = \rightarrow_{\text{local}}(\mathcal{S}(\mathcal{A}))$ for any $\mathcal{A}$?

**Theorem 5.15**

Splittability for disjoint splitters and regular document spanners is PSPACE-complete.
Split-Constrained Black Boxes

- Information Extraction in practice is often done with other algorithms, e.g.
  - coreference resolvers
  - sentiment extractors
  - named entity recognition

Split-Correctness

Input: Black Box Extractors $\rightarrow_1$, $\rightarrow_2$, a Splitter $\blacklozenge$, and a set $C$ of Split-Constraints

Question: Is $\rightarrow_1 \bowtie \rightarrow_2$ splittable by $\blacklozenge$ under $C$?
Split-Constrained Black Boxes

**Black Box Split-Correctness**

**Input:** Black Box Extractors \(\rightarrow_1, \rightarrow_2\), a Splitter \(\mathcal{K}\), and a set \(C\) of Split-Constraints

**Question:** Is \(\rightarrow_1 \bowtie \rightarrow_2\) splittable by \(\mathcal{K}\) under \(C\)?

There are cases, where \(\rightarrow_1 \bowtie \rightarrow_2\) is splittable by sentences, even though \(\rightarrow_1\) is not!

All 'lost' tuples of \(\rightarrow_1\) are not needed for the join

**Theorem 7.3**

There are extractors \(\rightarrow_1\) and \(\rightarrow_2\) that are self-splittable by the same splitter \(\mathcal{K}\)

but \(\rightarrow_1 \bowtie \rightarrow_2\) is not splittable by \(\mathcal{K}\) under \(C\)
Why should I care about this?

Main Motivation

Parallelized information extraction for huge documents

Further Motivation

1. Reduce the skew, if the a big number of smaller documents must be analyzed
2. Debugging: notify users if extractors reach beyond natural boundaries
3. Handling updates to the input document
Future Work

- Complexity of splittability for non-disjoint splitters
- Try to better understand splittability of split-constrained black boxes
- Empirical analysis of the framework

Thank you for your attention 😊

Paper: doi.org/10.1145/3294052.3319684