Tackling the “We have no Data” Challenge: Domain-Specific Machine Translation in SMEs

AI-DRSWA: Maturing Artificial Intelligence - Data Science for Real-World Applications

Dr. Frederik Simon Bäumer
Bielefeld University of Applied Sciences

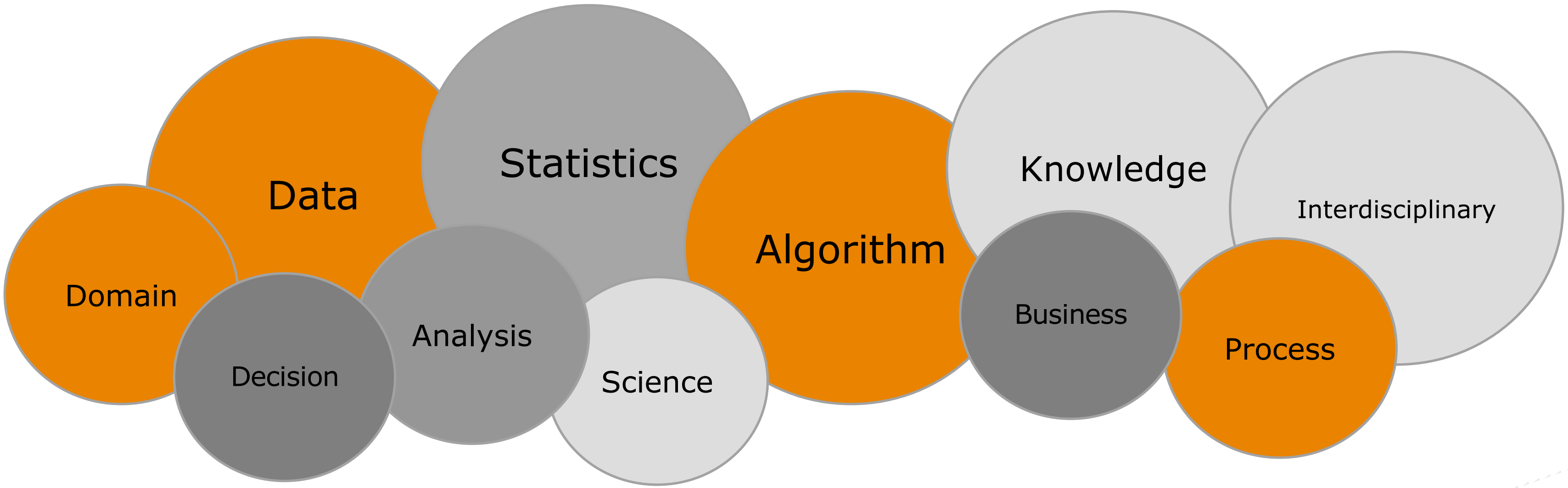
Practical report in cooperation with Wonki GmbH, Bielefeld.

Data Science View

Definition: The bad news

The situation

- No unambiguous, universally accepted definition of Data science
- *"One of the most generic and vaguely defined fields of study to have come about in the past 50 years"* (Peng & Parker, 2021)
- Several definitions exist; share some terms:

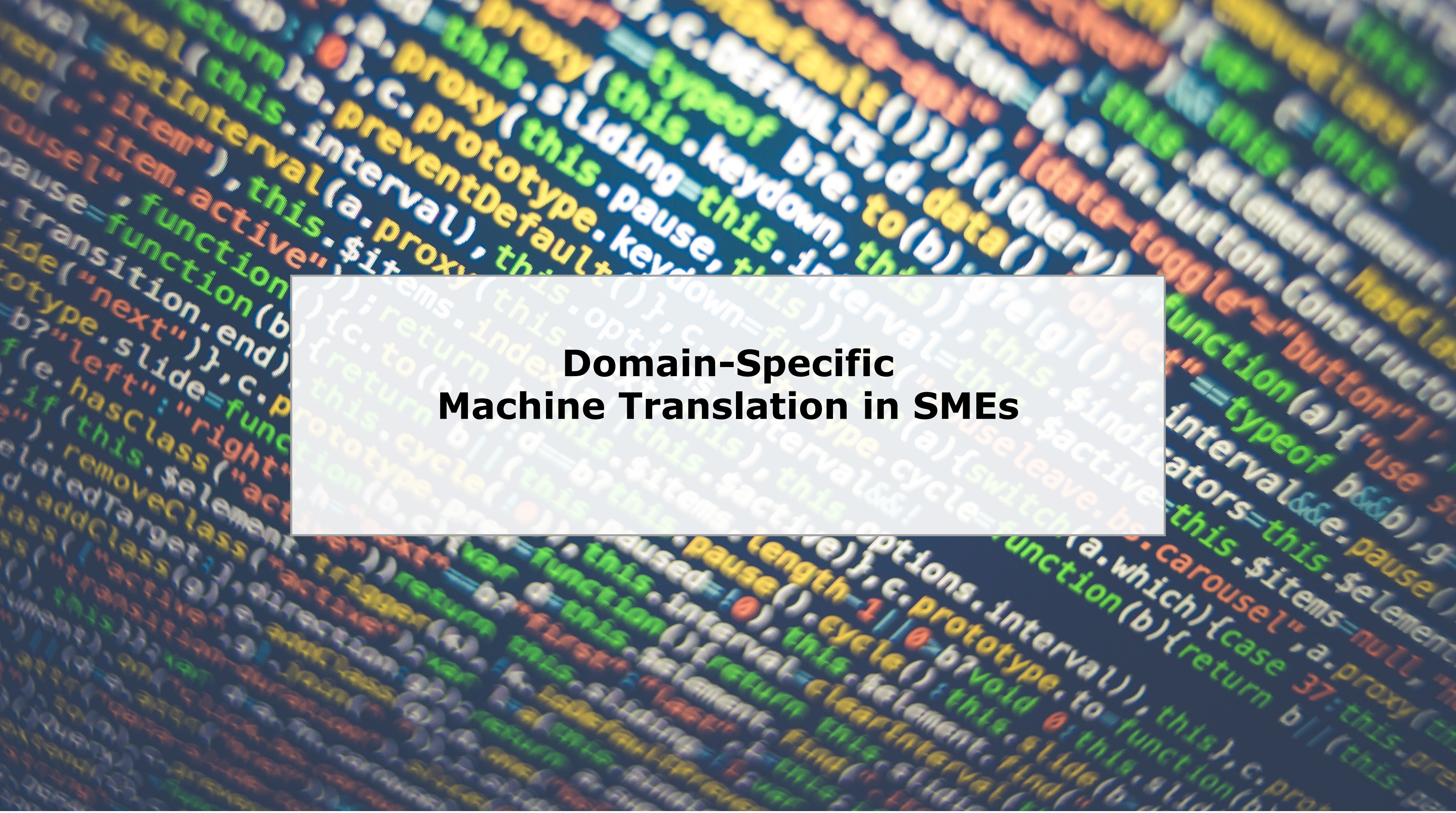


(Peng & Parker, 2021; Maslianko & Sielskyi, 2021; Van Der Aalst, 2016)

Data Science Definition

“Data Science is an **interdisciplinary field** that uses scientific methods, processes and algorithms to extract **insights** from structured and **unstructured data** to **systematically** achieve **competitive advantage**”

(AG Angewandte KI @ FH Bielefeld; GI, Working Paper, 2019; Peng & Parker, 2021)



Domain-Specific
Machine Translation in SMEs

Why Machine Translation?

- **Translation software has advantages for companies**
  - Facilitate communication
  - Facilitate editing and creation of multilingual documents
  - Results are immediately available
  - Results can be adapted flexibly
- **Nevertheless, concerns exist:**
  - Translation quality
  - Specialized vocabulary
  - Industry-specific phrases,
  - Data security

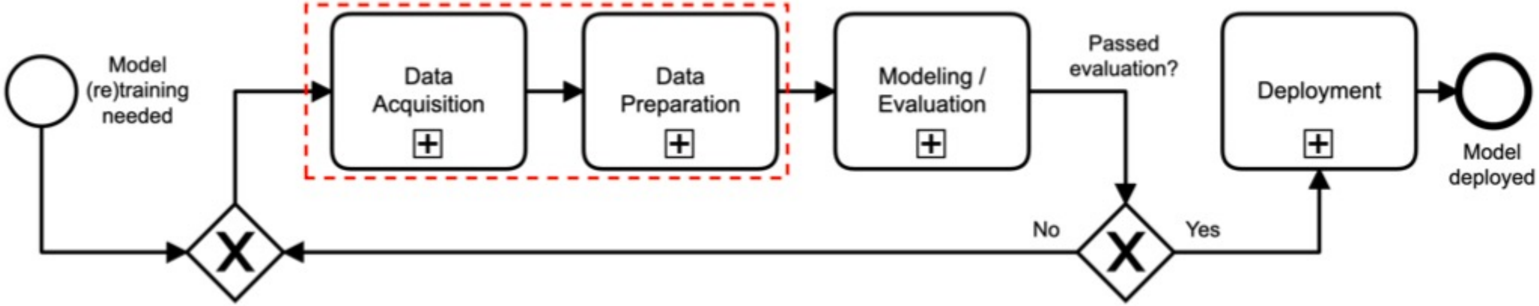
Why Machine Translation?

- **Self-hosted business-specific translation models can address problems**
  - Increasing speed
  - Providing company-specific translations
  - Data security
- **However: Companies need to contribute necessary training data**
  - Companies are sitting on a treasure of data that needs to be lifted
  - Processes and software to create datasets for translation solutions needed
  - Company-specific workflows needed
  - Human-in-the-loop tools for translation quality review

Workflows

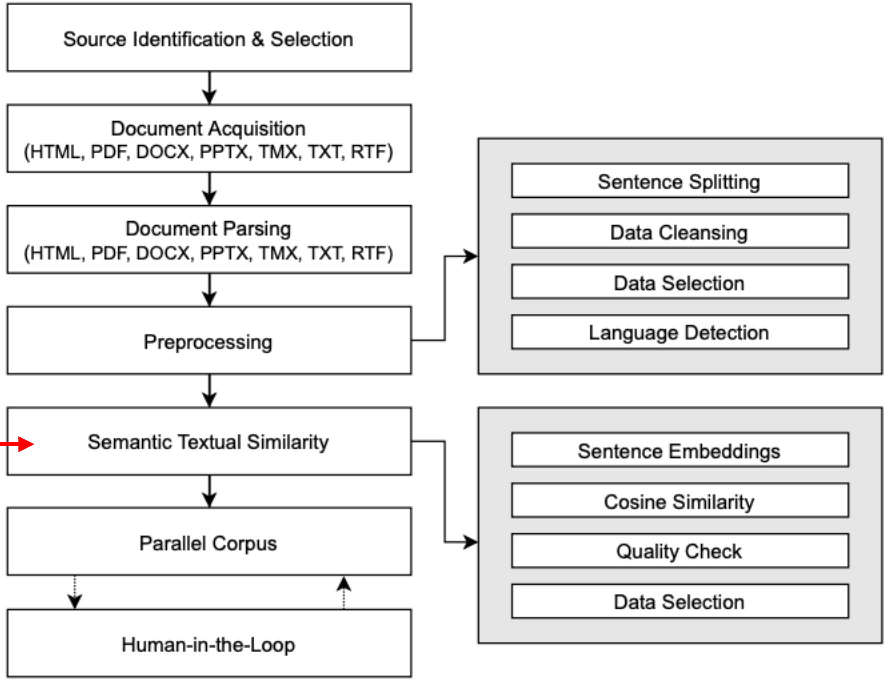
Overall Training Process

- CRISP inspired workflow
- From data acquisition to model deployment
- We need to support companies in each step

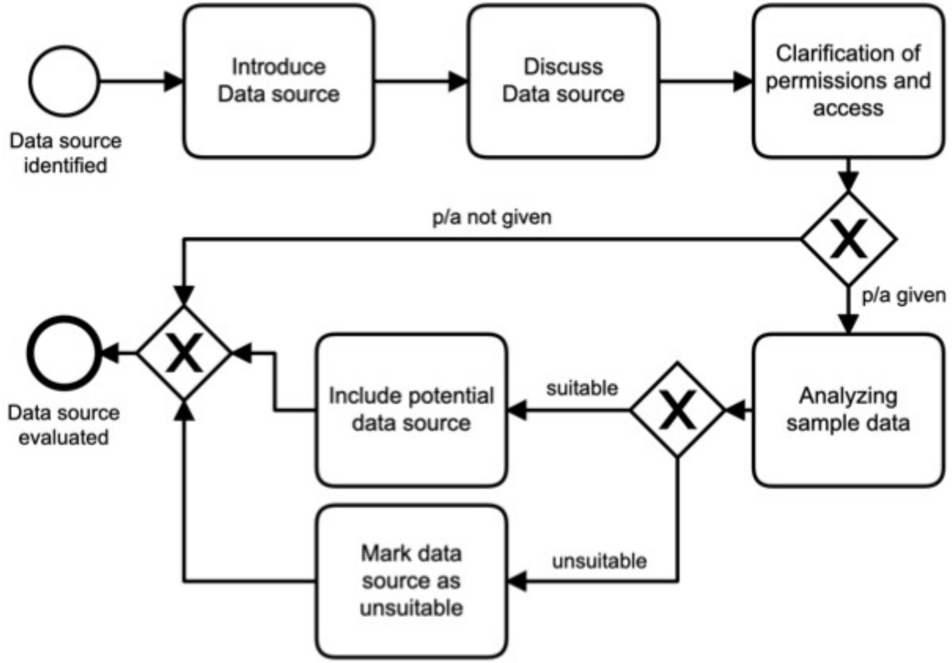


Data Acquisition & Processing

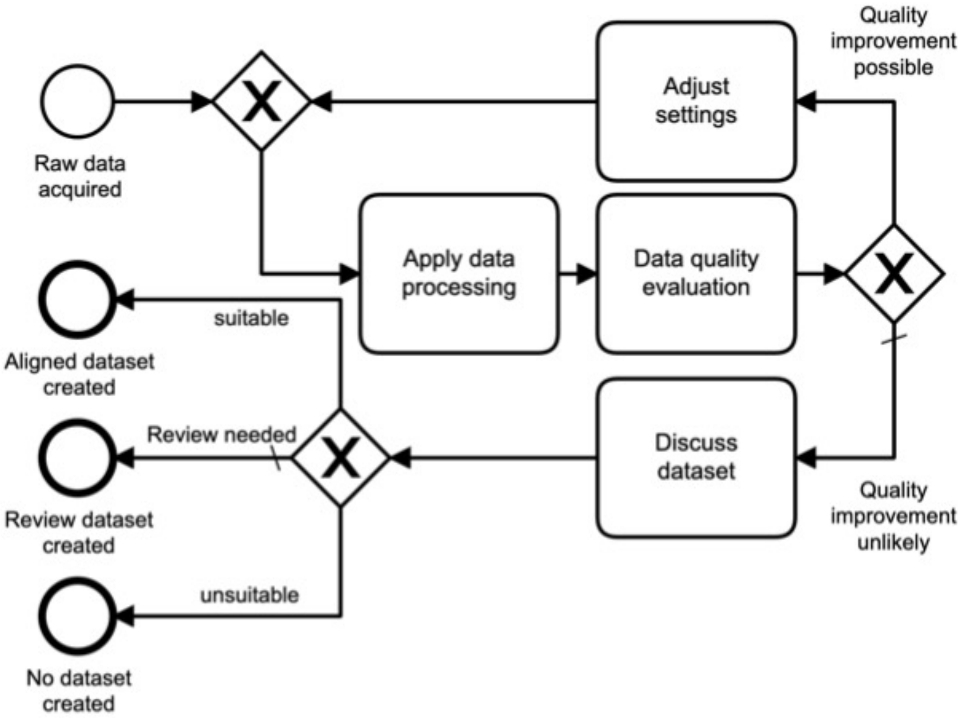
LaBSE model supports 109 languages


Source Identification & Selection
Document Acquisition (HTML, PDF, DOCX, PPTX, TMX, TXT, RTF)
Document Parsing (HTML, PDF, DOCX, PPTX, TMX, TXT, RTF)
Preprocessing
Semantic Textual Similarity
Parallel Corpus
Human-in-the-Loop
Sentence Splitting
Data Cleansing
Data Selection
Language Detection
Sentence Embeddings
Cosine Similarity
Quality Check
Data Selection

Data Source Identification



Dataset Preparation



Learnings

- There is **skepticism** about MTs
- Convince and involve **people**
- Communicate that any form of text data is interesting in the first place
- **Automate** repetitive, time-consuming tasks
- Established **evaluation tools** and support the users
- **Human-In-The-Loop** is a key success factor

Conclusion

- In MT for SME: The **question of suitable training data** quickly arises
- Data is spread across **various** platforms, cloud storage, and systems
- We support companies in finding **relevant** datasets and preparing them
- We support **fine-tuning** translation models
- Lack of understanding of how data can **contribute** to a translation system
- We established **evaluation tools** and support the users
- **Human-In-The-Loop** is a key success factor

Thank you