Graduate School of Excellence advanced Manufacturing Engineering in Stuttgart

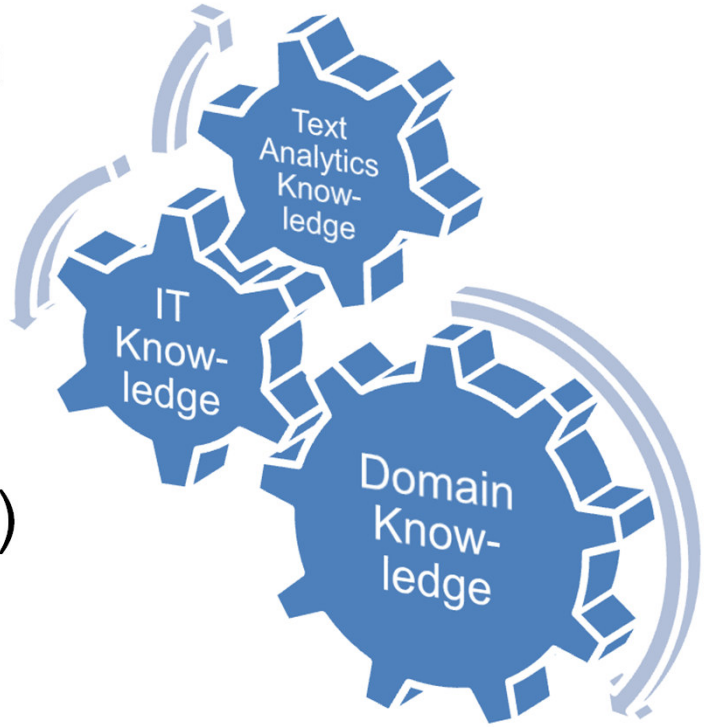
Die Gratwanderung zwischen qualitativ hochwertigen und einfach zu erstellenden domänenspezifischen Textanalysen

(High quality and easy domain specific text analysis: A balancing act)

Cornelia Kiefer
Graduate School of Excellence advanced Manufacturing Engineering
Universität Stuttgart

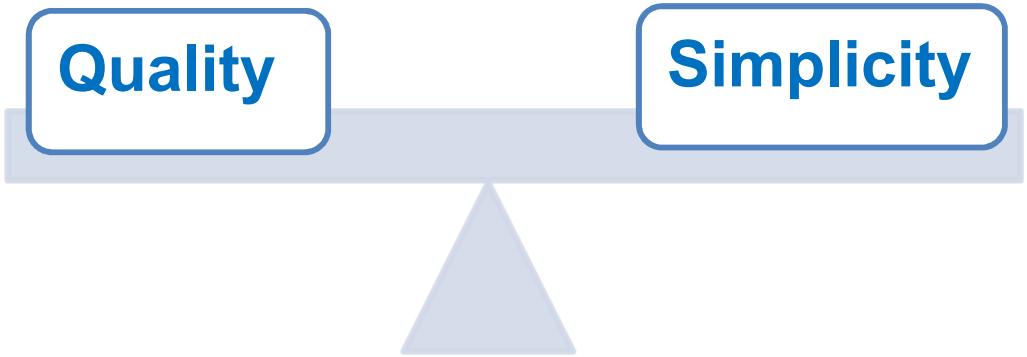
Introduction

- Text Analytics answers questions in humanities, sciences and industry
- IT, Text Analytics and Domain Knowledge needed
- Solution: simplified ad hoc Text Analysis (e.g. Leipzig Corpus Miner)



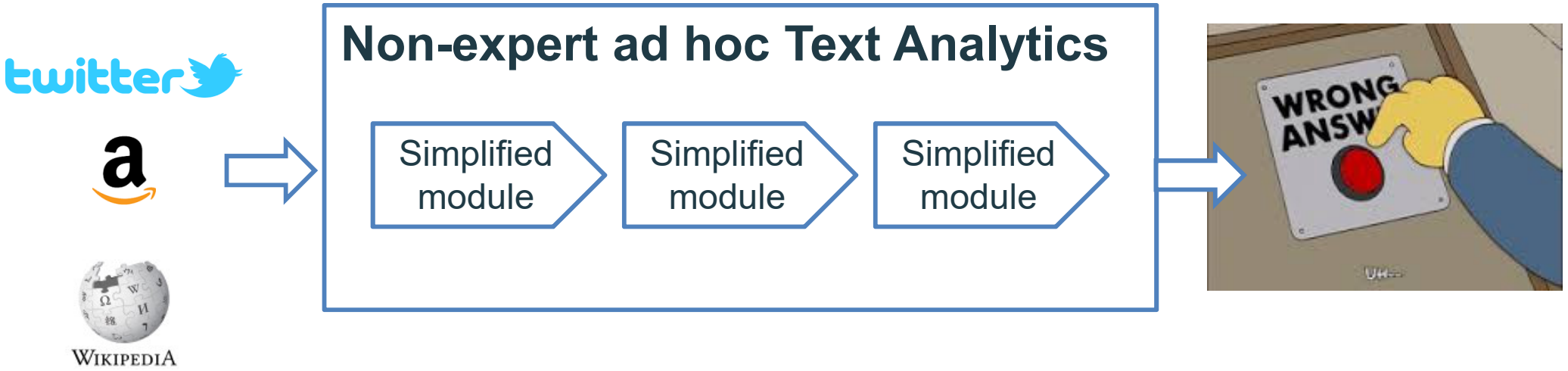
Introduction

- But how to ensure that:
  - simplified != less quality
  - Ordinary user (domain expert) of Text Analytics gets high quality results
- This work: Data Quality
  - Illustrate two data quality indicators and reveal quality problems in simplified text analytic pipelines
  - Basis for:
    - Consumer-oriented Text Analytics
    - Improvement of simplified ad hoc Text Analytics



Motivation

- A simplified analysis of bad quality text data can lead to wrong research results in humanities and sciences
- A simplified analysis of various text types with a simplified consumer which is expecting only one type of text leads to bad quality results



Bilderquellen: http://www.simpsonsworld.com, https://www.amazon.de/, https://twitter.com, https://www.wikipedia.de/

Research Question

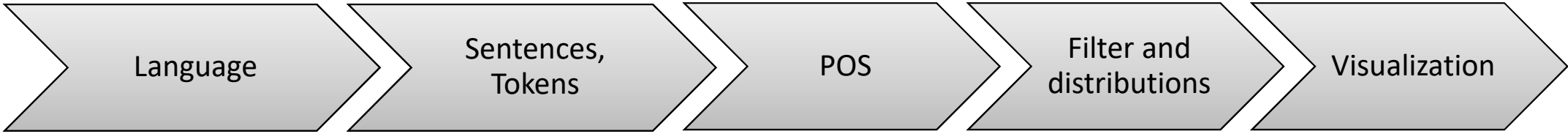
- Illustrate two data quality indicators in a use case from humanities
- Show problems of simplifications of non-expert text analytics and modules therein

Related work

- **Little research on data quality of unstructured text data**
- [So04] lists categories for data quality dimensions for unstructured text data together with indicators
- In [Ki16] we suggested data quality dimensions for unstructured data and listed concrete indicators for text data, two of them are illustrated in this work on a use case in humanities
- **In [So04] and [Ki16] data quality problems are not illustrated with real data and in a use case scenario as in this work**

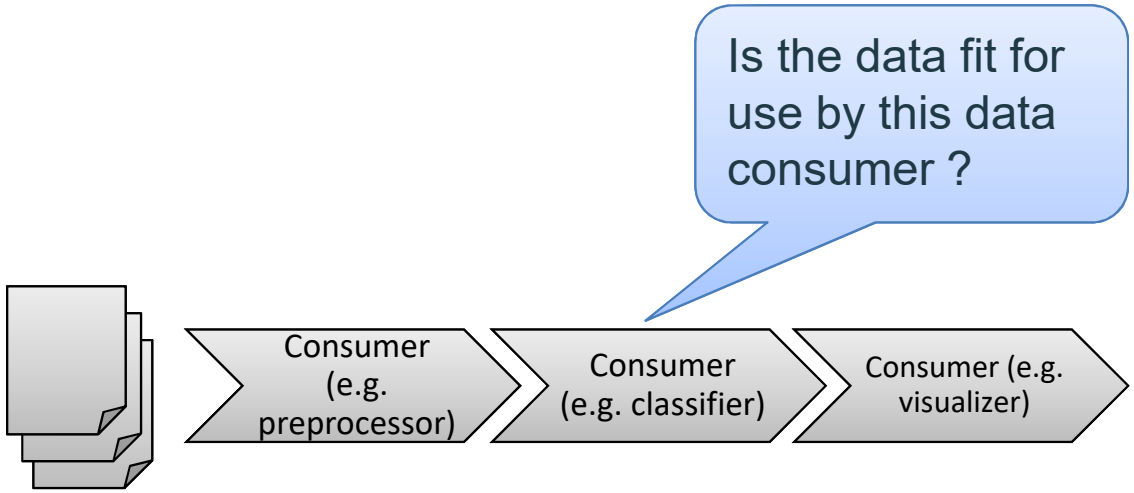
Use Case

- A linguist wants to analyse young people‘s language in social networks
- E.g. via distribution of adjectives or nouns in selected social media data over time
- Task leads to the text analysis pipeline below



Data Quality

- „Fitness for use by the data consumer“ [WS96]
- Algorithms and Text Analytic Modules are also data consumers:

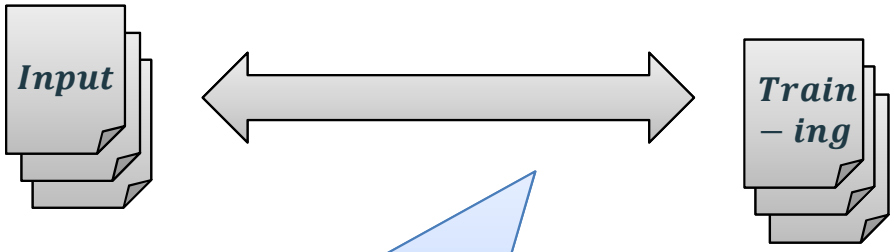


Indicators to Measure the quality of text data

Indicators [Ki16]:

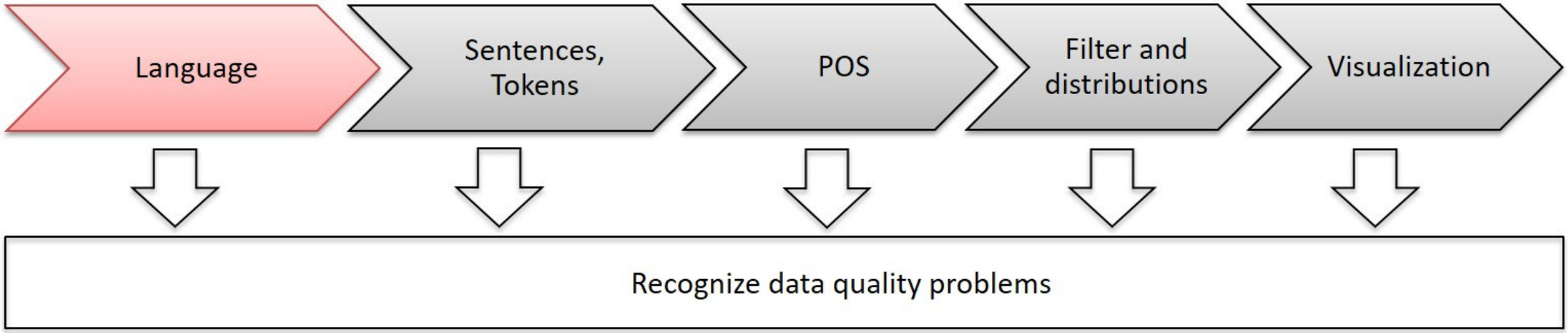
- Percentage of noisy data:
  - spelling mistakes
  - Abbreviations
  - unknown words

- Fit of training data:



Fit of training data: **Similarity** between input data and the data fit for use by a classifier, represented by it‘s training data

Identification of DQ problems in the use case



Language Identification

Data	Accuracy		
	Tika[1]	Language-detector[2]	Language-Identifier[3]
News (Penn Treebank, see [MMS93])	86	96	96
Novels (Brown, see [FK79])	84	89	91
Tweets (Twitter corpus, see [De13])	47	72	77
Chat posts (NPS Chat, see [FLM])	20	22	33

Information on data consumers:

- **Tika** trained on clean data → expects **clean** data
- **Language-detector** was also trained on tweets → expects **clean data and tweets**
- **LanguageIdentifier** trained on newsgroups → expects **newsgroups** texts

Examples for wrong classified texts:

„where did everyone gooo ?"

„ah well"

Maybe a good predictor: Percentage of noise

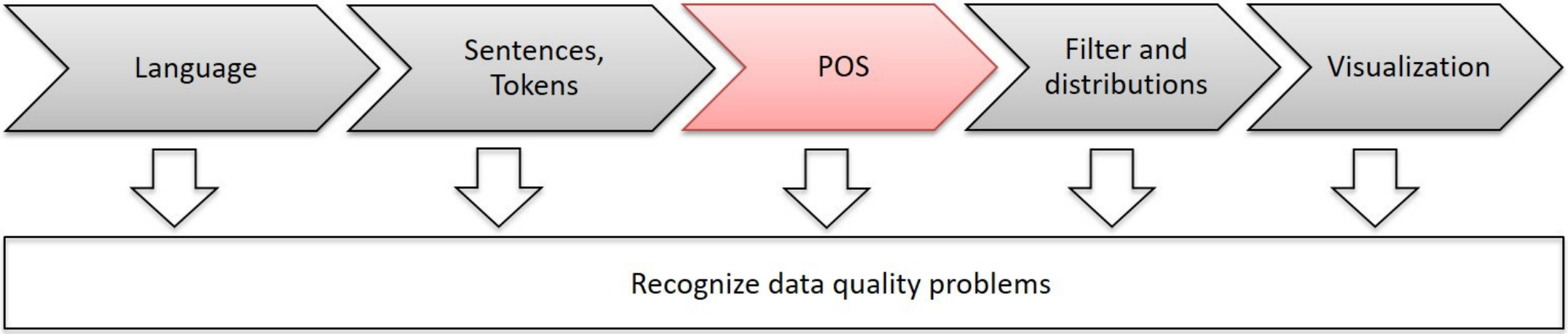
"RT @xochinadoll: I fel so blah today."

1) https://tika.apache.org/

2) https://github.com/optimaize/language-detector

3) DKPro Core: https://dkpro.github.io/dkpro-core/

Identification of DQ problems in the use case



Part of Speech Tagging

e.g. used in Leipzig Corpus Miner

Data	Accuracy		
	NLTK[1]	Stanford[2]	OpenNLP[3]
News (Penn Treebank, see [MMS93])	100	91	90
Novels (Brown, see [FK79])	60	63	63
Tweets (Twitter corpus, see [De13])	65	67	70
Chat posts (NPS Chat, see [FLM])	64	62	62

Maybe a good predictor: Similarity between clean data and Chat posts?

Information on consumers:

- Standard tools **trained on clean data** (from left to right: Penn Treebank, Wall Street Journal, not specified) → expect clean data

1) http://www.nltk.org/, http://nlp.stanford.edu/software/tagger.shtml, https://opennlp.apache.org/

Conclusion and Future Work

- Illustration of two concrete problems in a non-expert text analysis pipeline
- These **problems might be automatically identified using the two suggested data quality indicators** (before executing the analysis pipeline, without annotated data)**:**
  - fit of training data
  - percentage of noisy data

Future Work:

- Implement the indicators and solutions deduced from these indicators
- **Integrate these solutions (e.g. automatic selection of best fitting training data, automatic correction of noisy data) to simplified ad hoc text analytics**

Contact Information

Cornelia Kiefer
Cornelia.Kiefer@gsame.uni-stuttgart.de

References

[De13] Derczynski, L. et al.: Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data: Proceedings of the International Conference on Recent Advances in Natural Language Processing. Association for Computational Linguistics, 2013.
[FK79] Francis, W. N.; Kučera, H.: Manual of Information to Accompany A Standard Corpus of Present-day Edited American English, for Use with Digital Computers. Brown University, Department of Lingustics, 1979.
[FLM] Forsyth, E.; Lin, J.; Martell, C.: The NPS Chat Corpus. http://faculty.nps.edu/cmartell/NPSChat.htm, 03.11.2016.
[Ki16] Kiefer, Cornelia (2016): Assessing the Quality of Unstructured Data: An Initial Overview. In: Ralf Krestel, Davide Mottin und Emmanuel Müller (Hg.): CEUR Workshop Proceedings. Proceedings of the LWDA. Potsdam. Aachen (CEUR Workshop Proceedings), S. 62–73. Online verfügbar unter http://ceur-ws.org/Vol-1670/#paper-25.
[MMS93] Marcus, M. P.; Marcinkiewicz, M. A.; Santorini, B.: Building a Large Annotated Corpus of English: The Penn Treebank. In Comput. Linguist., 1993, 19; S. 313–330.
[So04] Sonntag, Daniel (2004): Assessing the Quality of Natural Language Text Data. In: GI Jahrestagung. 1. Aufl., S. 259–263.
[WS96] R. Y. Wang and D. M. Strong. Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst., 12(4):5{33, 1996.