

Extrahierung von Anforderungen aus natürlich-sprachlichen Lastenheften: Was erschwert eine KI-basierte Extrahierung?

Prof. Dr.-Ing. Iris Gräßler

Daniel Preuß

Dr.-Ing. Jens Pottebaum

Universität Paderborn, Heinz Nixdorf Institut, Fürstenallee 11, 33102 Paderborn,
{iris.graessler, daniel.preuss, jens.pottebaum}@hni.upb.de

1. Motivation und Problemstellung

Bei der Entwicklung komplexer technischer Systeme werden ca. 80 % der Benutzeranforderungen in natürlich-sprachlichen Anforderungsdokumenten dokumentiert [1]. Das Lastenheft für ein Flugzeugtriebwerk von Rolls-Royce umfasst beispielsweise über 700 Benutzeranforderungen, wodurch der Aufwand für die Extrahierung dieser Anforderungen hoch ist [2]. Die manuelle Durchführung dieser Aktivität ist zeitaufwändig und monoton. Daraus folgen Fehleranfälligkeit und hohe Kosten für Unternehmen. Darum ist in diesem Anwendungskontext eine automatisierte Extrahierung notwendig. In der Software-Entwicklung werden Techniken des Natural Language Processing (NLP) für Requirements Engineering-Aufgaben angewendet.

Ansätze existieren auch für die KI-basierte Extrahierung von Anforderungen. Hintergrund des Beitrags ist unter anderem ein Ansatz, in dem Text-Segmentierung (vgl. [3]) und Klassifizierung durch Ansätze des maschinellen Lernens angewendet wurden (vgl. [1, 4]). Auf Basis eines Vergleichs mit LSTM- und BERT-Modellen wurde im Beispiel eine Support Vector Machine eingesetzt. Verschiedene Aspekte erschweren jedoch die Anwendung auf klassische Lastenhefte.

Ziel dieses Beitrags ist es, derartige Aspekte aufzuzeigen, die eine KI-basierte Extrahierung erschweren. Zukünftige Ansätze zur KI-basierten Extrahierung von Anforderungen aus Lastenheften können diese Aspekte systematisch adressieren.

2. Vorgehen

Die Datengrundlage bilden Beobachtungen aus Projekten, die z. B. in Form von Lastenheften, darin maschinell erfassten Text-Segmenten und der manuellen Verifikation dieser Text-Segmente vorliegen. Um beobachtete Aspekte abzusichern, wird eine Literaturanalyse durchgeführt (siehe Kapitel 3). Ansätze zur Extrahierung von Anforderungen werden dabei hinsichtlich ihrer Anwendbarkeit

auf Lastenhefte von zwei der größten Engineering-Dienstleistern aus dem Automotive-Bereich sowie Open Source Software-Spezifikationen analysiert. Einschränkende Aspekte werden systematisch erfasst (siehe Kapitel 4). Der Ansatz zur Extrahierung wird iterativ angepasst, um sämtliche Textsegmente verarbeiten zu können. Ursachen für die Adaption werden abstrahiert und als Aspekt deklariert. Fazit und Ausblick fassen Schlüsse aus den Ergebnissen zusammen (siehe Kapitel 5).

3. Ansätze zur KI-basierten Extrahierung von Anforderungen aus Lastenheften

KURTANOVIC und MAALEJ wenden überwacht maschinelles Lernen an, um Sätze einer Software-Spezifikation zu klassifizieren [5]. Merkmale, die für die Klassifizierung genutzt werden, sind beispielsweise N-Gramme der Wörter und der Part-of-speech-Tags (POS-Tags). HARIS und KURNIAWAN definieren semantische Regeln, die auf POS-Tags basieren, um natürlich-sprachliche Anforderungen zu identifizieren, die einem vordefinierten Schema folgen [4]. Solche Anforderungsschablonen (zum Beispiel nach [6]) umfassen in Abhängigkeit der Anforderungsart die jeweiligen Bestandteile einer Anforderung – zum Beispiel Akteur und Prozesswort. METH, MUELLER und MAEDCHE verarbeiten natürlich-sprachliche Rezensionen von Nutzern auf Online-Marktplätzen, um Anforderungen für eine neue Produktgeneration zu ermitteln [1].

Die Ansätze haben gemein, dass zunächst eine Text-Segmentierung durchgeführt werden muss. Hierbei werden z. B. Co-Referenzen aufgelöst. Dies sind Anforderungen mit Bezug zu Objekten, die in einem vorherigen Segment beschrieben wurden. Im nächsten Schritt werden die Text-Segmente klassifiziert: *funktionale Anforderung*, *nicht-funktionale Anforderung* oder *keine Anforderung*. Abschließend muss die Anforderungs-Qualität geprüft werden (siehe Abbildung 1).

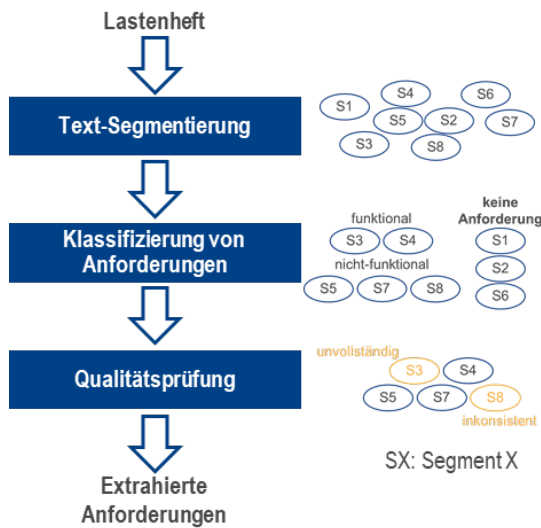


Abbildung 1: Extrahierung von Anforderungen

4. Beobachtungen: Was erschwert eine KI-basierte Extrahierung?

Das *Dateiformat* des Lastenhefts muss für die Extrahierung berücksichtigt werden. Beispielsweise werden in .docx-Dokumenten bestimmte Word-Vorlagen genutzt, um Textelemente auszuzeichnen. In der Text-Segmentierung müssen die entsprechenden Informationen verarbeitet werden. Die Bildung von Textsegmenten wird erschwert aufgrund verschiedener *Formatierungen im Text*: beispielsweise Aufzählungen, die auf einen einleitenden Satz aufbauen. Im Lastenheft werden *Verweise* auf Richtlinien und Normen gesetzt (Querverweise). Co-Referenzen müssen für die Bildung von Textsegmenten aufgelöst werden.

Anforderungen können *Qualitätsdefizite* enthalten. Unvollständige oder mehrdeutige Anforderungen führen zu Klassifizierungsfehlern. *Anforderungsbeschreibungen* sind teilweise unstrukturiert. Ein einheitliches Formulierungsschema z. B. in Form eines standardisierten Satzbaus kann im Gegensatz zu strukturierten Anforderungsdokumenten nicht vorausgesetzt werden. Außerdem existieren Unterschiede in der Formulierung verschiedener *Anforderungsarten* (funktionale und nicht-funktionale Anforderungen). Die Datengrundlage muss ausreichend groß sein, damit das Modell verschiedene Formulierungen korrekt klassifizieren kann. Zusätzlich müssen ausreichend viele fachspezifische Daten in den jeweiligen *Sprachen* verfügbar sein. Die Aspekte, beispielhafte Ausprägungen und Herausforderungen werden in der folgenden Tabelle 1 aufgeführt.

Aspekt	Ausprägung	Herausforderung
Dateiformat	.docx, .pdf, .xlsx	Textelemente erkennen
Formatierung im Text	Aufzählung, Tabelle	Bildung von Textsegmenten
Verweise	Querverweise, Co-Referenzen	Bildung von Textsegmenten
Qualitätsdefizite	unvollständig, mehrdeutig	unvollständige Merkmale
Anforderungsbeschreibung	unstrukturiert, strukturiert	keine konsistenten Regeln
Anforderungsarten	funktional, nicht-funktional	Multiklassen-Klassifikation
Sprache	Deutsch, Englisch	Datengrundlage für mehrere Sprachen

Tabelle 1: Aspekte, die eine KI-basierte Extrahierung erschweren

5. Fazit und Ausblick

Die automatisierte Extrahierung von Anforderungen wird durch verschiedene Aspekte erschwert. Lastenhefte sind nicht standardisiert, weshalb eine intelligente Verarbeitung der Lastenhefte durch KI-Ansätze unter Berücksichtigung der aufgeführten Aspekte benötigt wird.

6. Literatur

- [1] Meth, H.; Mueller, B.; Maedche, A.: "Designing a Requirement Mining System". In: Journal of the Association for Information Systems, Jg. 16, Nr. 9. 2015, S. 799-837.
- [2] Fernandes, J.; Henriques, E.; Silva, A.; Moss, M. A.: "Requirements change in complex technical systems: an empirical study of root causes". In: Research in Engineering Design, Jg. 26, Nr. 1. Springer London, 2015, S. 37-55.
- [3] Gräßler, I.; Preuß, D.: "Anwendbarkeit von Requirement Mining in Benutzerrezensionen für die Entwicklung mechatronischer Produkte im B2C-Markt". In: Digital-Fachtagung Mechatronik, Darmstadt, 2021 S. 68-73, 2021
- [4] Haris, M. S.; Kurniawan, T. A.: "Automated requirement sentences extraction from software requirement specification document". In: Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology, NY, 2020, S. 142-147.
- [5] Kurtanovic, Z.; Maalej, W.: "Automatically Classifying Functional and Non-functional Requirements Using Supervised Machine Learning". In: 25th International Requirements Engineering Conference, 2017, S. 490-495.
- [6] Rupp, C.: Requirements-Engineering und -Management. Hanser, 2014, S. 215-246