

Identifikation von Rückverfolgbarkeitsverbindungen zwischen Anforderungen mittels Sprachmodellen

Dokumentenart: Exposé für eine Bachelorarbeit
Autor: Niklas Ewald
Matrikel-Nr.: 1996334
Studiengang: Informatik Bachelor
Betreuer: Tobias Hey
Datum: 15. Dezember 2020

1 Motivation

Die Rückverfolgbarkeit von Anforderungen ist ein wichtiger Teil der Softwareentwicklung. Sie erlaubt es unter anderem zu belegen, dass alle Anforderungen des Auftraggebers umgesetzt wurden. Außerdem können Rückverfolgbarkeitsinformationen verwendet werden um zu überprüfen, welche Auswirkungen eine Änderung auf andere Teile des Projekts hat [CHGZ12]. In einigen sicherheitsrelevanten Bereichen, wie zum Beispiel in der Luftfahrt, wird die Rückverfolgung von Anforderungen sogar explizit gefordert. Manuelle Rückverfolgung von Anforderungen ist zeitaufwändig und anfällig für Fehler. Automatisierte Verfahren verwenden Informationsrückgewinnungsmethoden oder maschinelles Lernen.

Ein Problem beim Generieren von Rückverfolgbarkeitsverbindungen zwischen natürlichsprachlichen Anforderungen sind die unterschiedlichen Abstraktionslevel der Anforderungen, die die exakte Bestimmung semantischer Ähnlichkeit erschweren.

Eine Möglichkeit, um dieses Problem zu lösen, könnte die Verwendung von Sprachmodellen wie BERT [DCLT19] sein. Diese Modelle werden auf großen Sprachkorpora vortrainiert, um generelle Eigenschaften von natürlicher Sprache zu lernen. Durch das Training im Kontext eines Satzes werden Informationen über Beziehungen zwischen Wörtern gelernt. Das Verständnis der Eigenschaften von natürlicher Sprache und das Verständnis der Bedeutung von Wörtern im Kontext könnten helfen, Ähnlichkeiten zwischen den Anforderungen zu finden, die mit anderen Methoden bisher nicht entdeckt werden. Abbildung 1 zeigt drei Anforderungen aus dem GANTT-Datensatz [HHD09].

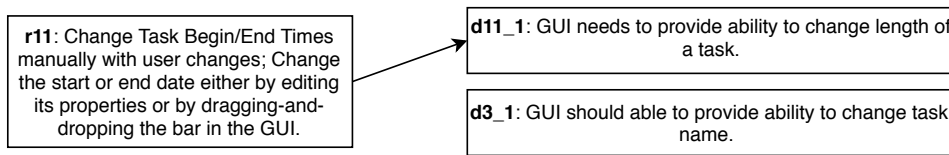


Abbildung 1: Beispiel für Anforderungen aus dem GANTT-Datensatz [HHD09] mit entsprechender Rückverfolgbarkeitsverbindung

Die Anforderungen d11_1 und d3_1 sind sehr ähnlich aufgebaut. Beide beschreiben die Möglichkeit ein Attribut über die Benutzeroberfläche zu ändern. Anforderung r11 betrifft die Änderung von Start- und Endzeiten von Aufgaben. Dies muss mit der Länge von Aufgaben in d11_1 in einen Zusammenhang gebracht werden. Möglicherweise haben Sprachmodelle durch das Training die semantische Verbindung zwischen diesen Begriffen gelernt, sodass eine Verbindung zwischen den Anforderungen r11 und d11_1, aber nicht zwischen r11 und d3_1 identifiziert werden kann.

2 Zielsetzung

Das Ziel dieser Arbeit ist der Entwurf eines Verfahrens zur automatischen Identifikation von Rückverfolgbarkeitsverbindungen zwischen natürlichsprachlichen Anforderungen. Zu diesem Zweck soll der Einsatz von Sprachmodellen zur Generierung von Rückverfolgbarkeitsverbindungen zwischen Anforderungen mit unterschiedlichen Abstraktionsgraden untersucht werden.

Hierfür müssen zunächst vorhandene Sprachmodelle untersucht werden, um herauszufinden, ob sie für die Lösung dieses Problems geeignet sind. Daraufhin soll ein geeignetes Fein Anpassungsverfahren der Sprachmodelle gefunden werden.

Abschließend soll ein Verfahren zur Abbildung der Anforderungen aufeinander entworfen werden. Bei der Entwicklung des Gesamtverfahrens soll außerdem darauf geachtet werden, dass es möglichst projektunabhängig ist und auch für Projekte, für die keine Trainingsdaten vorliegen, verwendet werden kann.

3 Vorgehen

Zunächst muss untersucht werden, welche vorhandenen Sprachmodelle infrage kommen, um Rückverfolgbarkeitsverbindungen zu generieren. Interessant sind vor allem Modelle, die bereits erfolgreich für ähnliche Aufgaben eingesetzt wurden, zum Beispiel für das Bestimmen der semantischen Nähe zwischen Sätzen in natürlicher Sprache. Für die gefundenen Sprachmodelle

muss überprüft werden, welche Methoden und damit welche Arten der Fein-
anpassung geeignet sein könnten, um Rückverfolgbarkeitsverbindungen zu
finden.

Eine Möglichkeit, manche Sprachmodelle, darunter BERT [DCLT19], zur Ge-
nerierung von Rückverfolgbarkeitsverbindungen zu nutzen, könnte es sein, sie
als Klassifikator zu nutzen. Dafür müssten sie so feinangepasst werden, dass
sie für zwei Anforderungen als Eingabe vorhersagen, ob es sich um zusam-
mengehörige Anforderungen handelt.

Ein anderes Vorgehen könnte es sein, mithilfe von Sprachmodellen Satzein-
bettungen für die einzelnen Anforderungen zu berechnen, die mit einem Ähn-
lichkeitsmaß, wie zum Beispiel der Kosinus-Ähnlichkeit, verglichen werden
können. Hat ein Paar von Anforderungen eine hohe Ähnlichkeit, könnte eine
Rückverfolgbarkeitsverbindung zwischen den entsprechenden Anforderungen
identifiziert werden. Ein Modell, das auf diese Weise verwendet werden könn-
te, ist Sentence-BERT [RG19].

Abschließend sollen die Vorgehensweisen miteinander verglichen werden, um
diejenige zu finden, die am besten geeignet ist um Rückverfolgbarkeitsver-
bindungen zu identifizieren.

4 Evaluation

Es soll gezeigt werden, wie gut das entwickelte Verfahren Rückverfolgbar-
keitsverbindungen zwischen natürlichsprachlichen Anforderungen identifiziert.
Dazu können Datensätze (zum Beispiel Infusion Pump, GANTT und CM-
1) benutzt werden, die bereits in Veröffentlichungen zur Identifikation von
Rückverfolgbarkeitsverbindungen [SV20], [FH19] verwendet wurden. Falls
das verwendete Sprachmodell feinangepasst werden muss, wird das Verfahren
in einem Kreuzvalidierungsverfahren evaluiert. Das entsprechende Sprach-
modell wird mit Teilmengen der Datensätze feinangepasst und auf den ver-
bleibenden Teilen getestet. Kennzahlen, die bei der Identifikation von Rück-
verfolgbarkeitsverbindungen verwendet werden, sind unter anderem Präzi-
sion und Ausbeute und F-Maß beziehungsweise durchschnittliche Präzision
[SHC15]. Mithilfe dieser Kennzahlen ist ein Vergleich mit existierenden Ver-
fahren möglich.

Literatur

[CHGZ12] CLELAND-HUANG, Jane (Hrsg.) ; GOTEL, Orlena (Hrsg.) ;
ZISMAN, Andrea (Hrsg.): *Software and Systems Traceability*. London : Springer London, 2012. <http://dx.doi.org/10.1007/978-1-4471-2239-5>. <http://dx.doi.org/10.1007/978-1-4471-2239-5>. – ISBN 978-1-4471-2238-8 978-1-4471-2239-5

- [DCLT19] DEVLIN, Jacob ; CHANG, Ming-Wei ; LEE, Kenton ; TOUTANOVA, Kristina: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota : Association for Computational Linguistics, Juni 2019, 4171–4186
- [FH19] FARRAR, D. ; HAYES, J. H.: A Comparison of Stemming Techniques in Tracing. In: *2019 IEEE/ACM 10th International Symposium on Software and Systems Traceability (SST)*, 2019, S. 37–44. – ISSN: 2157-2194
- [HHD09] HOLBROOK, E. A. ; HAYES, J. H. ; DEKHTYAR, A.: Toward Automating Requirements Satisfaction Assessment. In: *2009 17th IEEE International Requirements Engineering Conference*, 2009, S. 149–158. – ISSN: 2332-6441
- [RG19] REIMERS, Nils ; GUREVYCH, Iryna: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 11 2019
- [SHC15] SHIN, Y. ; HAYES, J. H. ; CLELAND-HUANG, J.: Guidelines for Benchmarking Automated Software Traceability Techniques. In: *2015 IEEE/ACM 8th International Symposium on Software and Systems Traceability*, 2015, S. 61–67
- [SV20] SCHLUTTER, A. ; VOGELSANG, A.: Trace Link Recovery using Semantic Relation Graphs and Spreading Activation. In: *2020 IEEE 28th International Requirements Engineering Conference (RE)*, 2020, S. 20–31. – ISSN: 2332-6441