Worteinbettungen für die Anforderungsdomäne

Dokumentenart:	Exposé für eine Bachelorarbeit
Autor:	Tobias Telge
Matrikel-Nr.:	2059534
Studiengang:	Informatik Bachelor
Betreuer:	Tobias Hey
Datum:	9. Dezember 2019

1 Motivation

In der Verarbeitung natürlicher Sprache (engl. *natural language processing*) spielen Worteinbettungen, die Wörter auf Wortvektoren abbilden, eine immer größere Rolle. Worteinbettungen berücksichtigen den Kontext, in dem Wörter vorkommen. So sollen Wörter mit ähnlicher Bedeutung auf ähnliche Wortvektoren abgebildet werden, wie in Abbildung 1 schematisch dargestellt. Es gibt verschiedene Verfahren und vortrainierte Worteinbettungsmodelle. Die Anwendung dieser vortrainierten Modelle auf Anforderungen kann jedoch zu Problemen führen. So könnte es zum Beispiel passieren, dass Wörter, die dem Modell nicht bekannt sind, auf denselben Wortvektor abgebildet werden, was nicht besonders hilfreich ist. Die Ursache für die Probleme ist, dass die vortrainierten Worteinbettungsmodelle auf Textkorpora trainiert wurden, in denen Begriffe aus der Anforderungsdomäne selten vorkommen, wie zum Beispiel Nachrichten.

Es besteht aber durchaus Bedarf an Worteinbettungen für die Anforderungsdomäne. Sie könnten zum Beispiel bei der Erkennung von äquivalenten Anforderungen oder bei der Klassifizierung von Anforderungen als funktional oder nichtfunktional helfen. Auch bei der Erkennung von nicht eindeutigen Formulierungen oder von Widersprüchen in Anforderungen könnten sie verwendet werden. Ein konkretes Beispiel ist das Projekt INDIRECT [Hey]. Hier wird zur Bestimmung der semantischen Funktion von Sätzen in Anforderungsbeschreibungen eines der vorhandenen Worteinbettungsmodelle verwendet. Das Ersetzen dieses durch ein für Anforderungen sinnvolleres Worteinbettungsmodell könnte dabei helfen, bessere Ergebnisse bei der Bestimmung der semantischen Funktion von Sätzen aus Anforderungen zu erzielen.

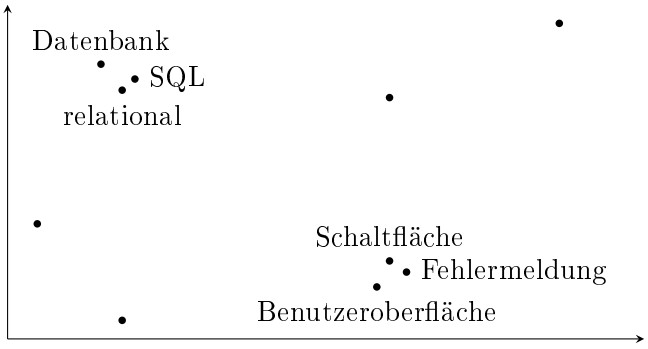

Abbildung 1: Schematische Darstellung von Worteinbettungen im Vektorraum

2 INDIRECT

In der Wartung und Pflege von Software hat die Rückverfolgbarkeit zwischen Quelltext und den entsprechenden Anforderungen eine große Bedeutung. Das Projekt INDIRECT (*Intent-driven Requirements-to-Code Traceability*) [Hey] hat zum Ziel, automatisiert Informationen zu dieser Rückverfolgbarkeit zu generieren. Das wird erreicht, indem für den Quelltext und für die Anforderungen je ein Graph modelliert wird. Die Graphen repräsentieren dabei die Absichten der einzelnen Anforderungen bzw. Quelltextteile. Zwischen den beiden Graphen wird dann eine Zuordnung gebildet.

3 Zielsetzung

Diese Arbeit hat das Ziel, Worteinbettungen für die Anforderungsdomäne zu bilden. Es soll ein Worteinbettungsmodell entstehen, das auf englischsprachigen Anforderungsdokumenten trainiert wurde. Dazu sollen Trainingsalgorithmen untersucht und ein geeigneter ausgewählt, sowie ein Korpus von Anforderungen gebildet werden. Es soll überprüft werden, wie das Worteinbettungsmodell in geeigneter Form mit bestehenden Modellen kombiniert werden kann. Das entstandene Worteinbettungsmodell soll abschließend analysiert werden. Dazu wird es auf Anforderungen angewandt und die Ergebnisse mit denen bestehender Modelle verglichen. Außerdem wird die Struktur des entstandenen Vektorraumes mit den Strukturen der anderen Wortvektorräumen verglichen.

4 Vorgehensweise

Die Qualität von Worteinbettungen hängt vom Trainingsalgorithmus und den Trainingsdaten ab. Zuerst sollen bestehende Algorithmen, wie zum Beispiel `word2vec` [MCCD, MSC+], `fastText` [BGJM] oder `GloVe` [PSM], mit

Augenmerk darauf untersucht werden, welche Vor- und Nachteile diese für die Anwendung auf Anforderungen haben.

Anschließend wird ein Korpus von Anforderungen und anderen in der Anforderungsdomäne üblichen Dokumenten, wie z. B. User Stories, gebildet, auf denen die Worteinbettungen trainiert werden können. Quantität ist hierbei wichtiger als Qualität, da auch die Anforderungen auf denen die Worteinbettungen später angewandt werden nicht notwendigerweise höchsten qualitativen Ansprüchen genügen. User Stories werden berücksichtigt, da in ihnen üblicherweise ebenfalls in Anforderungen vorkommende Formulierungen und Begriffe verwendet werden. Einen Startpunkt für die Sammlung geeigneter Anforderungen könnte der Datensatz von Ferrari et al. [FSG] darstellen.

Dann wird ein geeigneter Trainingsalgorithmus ausgewählt. Dabei werden die zuvor untersuchten Vor- und Nachteile für die Anwendung auf Anforderungen sowie Eigenschaften des Korpus, wie zum Beispiel die Quantität der Anforderungen, berücksichtigt. Um auch die auf umfangreichen Korpora trainierten bestehenden Modelle zu nutzen, soll zudem untersucht werden, wie das Worteinbettungsmodell in geeigneter Form mit diesen Modellen kombiniert werden kann.

Hierfür werden bestehende Worteinbettungsmodelle auf ihre Eignung untersucht und ausgewählt. Ein Startpunkt zur Kombination von Worteinbettungen könnte die von Kameswara Sarma et al. [KSLS] vorgestellte Methode sein. In ihr werden Worteinbettungen, die auf umfangreichen generischen Korpora trainiert wurden, mit domänenspezifischen Worteinbettungen kombiniert, indem die kanonische Korrelationsanalyse (engl. *canonical correlation analysis*) eingesetzt wird und die erhaltenen projizierten Worteinbettungen nach einem Optimierungsansatz linear kombiniert werden.

Abschließend wird das Worteinbettungsmodell auf dem Korpus trainiert.

5 Evaluation

Um das entstandene Worteinbettungsmodell zu evaluieren, wird es auf Anforderungen angewandt. Für ausgewählte Begriffe wird untersucht, welche Wörter diesen Begriffen im Modell semantisch am ähnlichsten sind. Dasselbe wird mit anderen Worteinbettungsmodellen durchgeführt und die Ergebnisse verglichen. Die Struktur des entstandenen Wortvektorraumes wird mit den Strukturen von anderen Wortvektorräumen, die zum Beispiel aus dem Training des Worteinbettungsmodells auf anderen Textkorpora entstanden, verglichen.

In *INDIRECT* sollen Worteinbettungen zur Bestimmung der semantischen Funktion von Sätzen in Anforderungsbeschreibungen verwendet werden. Die durch die Verwendung der domänenspezifischen Worteinbettungen entstandenen Ergebnisse werden mit den Ergebnissen verglichen, die durch die Verwendung anderer Worteinbettungen entstanden.

Literatur

[BGJM] Bojanowski, Piotr ; Grave, Edouard ; Joulin, Armand ; Mikolov, Tomas: Enriching Word Vectors with Subword Information. 5, S. 135–146. http://dx.doi.org/10.1162/tacl_a_00051. – DOI 10.1162/tacl_a_00051

[FSG] Ferrari, Alessio ; Spagnolo, Giorgio O. ; Gnesi, Stefania: PURE: A Dataset of Public Requirements Documents. In: *2017 IEEE 25th International Requirements Engineering Conference (RE)*, IEEE. – ISBN 978–1–5386–3191–1, S. 502–505

[Hey] Hey, Tobias: INDIRECT: Intent-Driven Requirements-to-Code Traceability. In: *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, IEEE. – ISBN 978–1–72811–764–5, S. 190–191

[KSLS] Kameswara Sarma, Prathusha ; Liang, Yingyu ; Sethares, Bill: Domain Adapted Word Embeddings for Improved Sentiment Classification. In: *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, Association for Computational Linguistics, S. 51–59

[MCCD] Mikolov, Tomas ; Corrado, G.s ; Chen, Kai ; Dean, Jeffrey: Efficient Estimation of Word Representations in Vector Space, S. 1–12

[MSC+] Mikolov, Tomas ; Sutskever, Ilya ; Chen, Kai ; Corrado, Greg ; Dean, Jeffrey: Distributed Representations of Words and Phrases and Their Compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, Curran Associates Inc. (NIPS'13), 3111–3119

[PSM] Pennington, Jeffrey ; Socher, Richard ; Manning, Christopher: Glove: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, S. 1532–1543