Multiwort-Bedeutungsauflösung für Anforderungen

Dokumentenart:	Exposé für eine Bachelorarbeit
Autor:	Thomas Bartel
Matrikel-Nr.:	1974965
Studiengang:	Informatik Bachelor
Betreuer:	Tobias Hey
Datum:	9. Dezember 2019

1 Motivation

Software gewinnt zunehmend an Bedeutung im alltäglichen Leben und findet Anwendung in fast allen Bereichen der Gesellschaft und Arbeitswelt. Die einwandfreie Funktion der genutzten Software ist deshalb eine Grundvoraussetzung. Damit dies gewährleistet werden kann, werden bereits verschiedene Techniken verwendet, um sowohl die Qualität des Codes zu verbessern, als auch die Wartung und Pflege zu vereinfachen. Eine Möglichkeit dies umzusetzen, ist die Verwendung von Rückverfolgbarkeitsinformationen zwischen Anforderungen und Quelltext innerhalb eines Softwareprojektes. Diese Informationen müssen derzeit jedoch noch mit hohem Aufwand manuell erstellt werden.

Damit dieser Prozess von einem System automatisiert werden kann, muss der Inhalt der Anforderungen zunächst von diesem System verstanden werden. Eine Grundvoraussetzung hierfür ist es, die Bedeutung der einzelnen Wörter der Anforderungen zu kennen. Diese Aufgabe übernehmen sogenannte Systeme zur Wortbedeutungsauflösung (engl. word sense disambiguation (WSD)), die Wörtern ihre im jeweiligen Kontext verwendete Bedeutung zuordnen. Viele dieser Systeme arbeiten allerdings nur auf einzelnen Wörtern. Die in englischer Sprache verfassten Anforderungen beinhalten jedoch häufig Fachbegriffe und zusammengesetzte Substantive (engl. compound nouns), die aus mehreren Wörtern bestehen. Neben weiteren Ausdrücken mit mehreren Wörtern werden diese in dieser Arbeit durch den Begriff Multiwort-Ausdruck (MWA) zusammengefasst [CEM+17].

Die Wortbedeutungen auf die die WSD-Systeme abbilden, werden aus Wissensquellen bezogen. Diese Wissensquellen sind jedoch unvollständig und hindern die WSD-Systeme in einigen Fällen daran, einem Wort die richtige Bedeutung zuzuordnen. Vor allem bei Fachbegriffen sind die passenden Definitionen für ihre Bedeutung nicht immer in den Wissensquellen vorhanden.

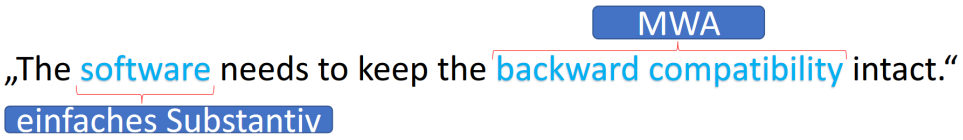

Abbildung 1: Anforderungsbeispiel mit MWA

Im Falle von MWAs kann man dennoch häufig Definitionen für die Teilworte in den Wissensquellen finden. Durch eine korrekte Bedeutungsauflösung bei MWAs in Anforderungstexten, würde die automatische Erzeugung von Rückverfolgbarkeitsinformationen erleichtert werden. Dies würde zur verbesserten Wartung und Pflege von Softwareprojekten führen. In Abbildung 1 ist ein Anforderungsbeispiel aufgeführt, in dem sowohl ein einfaches Substantiv, als auch ein MWA auftreten. Das Problem wird deutlich bei Betrachtung des MWAs „backward compatibility“. In der häufig verwendeten Wissensquelle WordNet [Wor] ist beispielsweise keine Definition für den Ausdruck zu finden, wodurch eine Bedeutungsauflösung schwierig wird. Für das Teilwort „compatibility“ sind jedoch zwei Definitionen vorhanden, von denen eine Rückschlüsse auf die Bedeutung des gesamten MWAs zulässt.

2 Projekt INDIRECT

Um die Wartung und Pflege von Software zu vereinfachen, entwickelt das IPD Tichy ein System, welches Rückverfolgbarkeitsinformationen zwischen Quelltext und Anforderungen automatisch erzeugt. Sowohl die Anforderungen in natürlicher Sprache, als auch der Quelltext werden hierbei in der Analyse miteinbezogen und in Form von Graphen modelliert [Hey19]. Für die Wörter der Anforderungen muss dafür eine Bedeutungsauflösung durchgeführt werden. INDIRECT arbeitet hierfür bereits mit WSD-Systemen, die einzelnen Wörtern eine Bedeutung zuordnen können. Dabei werden sowohl WordNet, als auch Wikipedia als Wissensquellen verwendet. MWAs werden hingegen noch nicht betrachtet. Eine vollständige Bedeutungsauflösung ist deshalb derzeit noch nicht in allen Fällen möglich.

3 Zielsetzung

Das Ziel dieser Bachelorarbeit ist der Entwurf und die Evaluation eines WSD-Systems, dem es möglich ist MWAs zu erkennen und ihnen eine korrekte Bedeutung in Abhängigkeit des gegebenen Kontexts zuzuordnen. Beim Zugriff auf Wissensbasen ist dabei zu beachten, dass in einigen Fällen nur für einzelne Bestandteile des MWAs eine passende Definition vorhanden ist. Das entwickelte Verfahren soll in diesem Fall im Rahmen der vorhandenen In-

formationen dennoch eine sinnvolle Bedeutungsauflösung durchführen. Der Fokus des entwickelten WSD-Systems soll, wie bereits in der Motivation erwähnt, auf Anforderungstexten liegen.

4 Vorgehensweise

Der erste wichtige Schritt ist zunächst die Bestimmung der möglichen MWAs innerhalb eines Satzes. Für ein WSD-System ist es nicht trivial zu erkennen, ob zwei Wörter eines Satzes zusammen einen neuen MWA mit eigener Bedeutung bilden. Arranz, Atserias und Castillo [AAC05] beschreiben ein Verfahren zur Bestimmung möglicher MWA. Hierzu verwenden sie eine geschlossene Liste solcher Ausdrücke aus WordNet und vergleichen mögliche Wortkombinationen innerhalb der Sätze mit dieser Liste. Dies ist nicht effektiv, da dieses Verfahren auf eine feste Liste beschränkt ist. Eine weitere Herangehensweise wäre die Untersuchung von Syntax - und Abhängigkeitsgraphen, um syntaktische Muster festzustellen, die bei der MWA-Detektion hilfreich sein könnten.

Nachdem automatisch alle zusammengesetzten Wörter in einem Satz gefunden wurden, muss diesen eine Bedeutung zugeordnet werden. Hierfür gibt es zum einen die wissensbasierten Ansätze, die mit manuell angelegten Lexika wie zum Beispiel WordNet arbeiten. Zum anderen existieren die korpusbasierten Ansätze, die mit realen Texten, wie zum Beispiel Wikipedia arbeiten. [McC09]. Außerdem treten auch Kombinationen dieser beiden Ansätze auf, wie man im Fall von Wikipedia und Wikidata sehen kann. Während Wikipedia die Informationen in Form von realen Texten präsentiert, agiert Wikidata als eine Wissensbasis.

Viele WSD-Systeme verwenden WordNet als primäre Wissensbasis für die Bedeutungsauflösung, da es bereits eine große Menge an Definitionen für allgemeine Kontexte bereitstellt. In dieser Arbeit soll das zu entwickelnde Verfahren jedoch mit domänenspezifischen Anforderungstexten arbeiten und häufig auch eine Bedeutungsauflösung bei Fachbegriffen durchführen. In diesem Fall bietet es sich eher an, eine stetig wachsende öffentliche Wissensquelle wie Wikipedia zu verwenden. Wie Rada Mihaelcea [Mih07] bereits demonstriert hat, lässt sich Wikipedia als mit Wortbedeutungen versehenes Korpus verwenden. Zunächst sollte also überprüft werden, ob ein oder mehrere Definitionen eines gefundenen MWAs in Wikipedia zu finden sind. Falls dies der Fall ist, wird eine Bedeutungsauflösung anhand der vorhandenen Definitionen durchgeführt.

Das beschriebene Vorgehen deckt die Fälle ab, in denen eine eindeutige Definition gefunden werden kann. Es besteht jedoch auch die Möglichkeit, dass weder in WordNet, noch in Wikipedia eine Definition für den vollständigen MWA vorhanden ist. In diesem Fall bietet es sich an die allgemeine Struktur von MWAs zu untersuchen, um herauszufinden, ob ein Zusammen-

hang zwischen der Bedeutung der Teilworte und der Bedeutung des gesamten Ausdrucks besteht. Ist dies der Fall, so sollte das Verfahren eine Bedeutungsauflösung bei den einzelnen Teilworten des MWAs durchführen. Anschließend wird der gefundene Zusammenhang genutzt, um Rückschlüsse auf die Bedeutung des gesamten Ausdrucks zu ziehen.

5 Evaluation

Um ein WSD-System zu evaluieren, muss getestet werden wie häufig das System eine korrekte Bedeutungsauflösung durchführt. Dazu benötigt man ein Textkorpus das bereits mit Bedeutungen annotiert wurde. Das zu testende WSD-System hat nun die Aufgabe, den Wörtern des Korpus eine Bedeutung zuzuordnen. Anschließend lassen sich verschiedene Metriken wie Präzision oder Ausbeute berechnen. Hierbei ist wichtig zu beachten, dass das zu entwickelnde WSD-System vor allem auf MWAs in Anforderungstexten ausgerichtet ist. Zur korrekten Evaluation sollten also auch Anforderungsspezifikationen verwendet werden. Da solche Texte im Normalfall noch nicht mit Bedeutungen annotiert wurden, muss dies zunächst manuell durchgeführt werden. Bei der Evaluation ist außerdem zu beachten, dass Definitionen der in den Dokumenten vorkommenden MWAs potenziell nicht in den verwendeten öffentlichen Wissensquellen vorhanden sind. Eine Bedeutungsauflösung mit der korrekten Definition ist in diesem Fall nicht möglich. Es besteht jedoch die Möglichkeit mehrere mögliche Bedeutungen auf eine gemeinsame Annotation abzubilden. Jede gefundene Bedeutung, die auf diese Annotation abgebildet wird, gilt dann als korrekt [PND07]. Mögliche sinnvolle Alternativbedeutungen die durch das WSD-System gefunden werden, können somit ebenfalls in die Evaluation einbezogen werden. Als abschließende Evaluation wäre noch ein Vergleich mit anderen WSD-Systemen, die ebenfalls für die MWA-Bedeutungsauflösung gedacht sind, möglich. Finlay und Kulkarni [FK11] haben ebenfalls solch ein System entwickelt. Ein Vergleich wäre möglich, in dem man dasselbe Textkorpus zur Evaluation verwendet.

Literatur

[AAC05] Arranz, Victoria ; Atserias, Jordi ; Castillo, Mauro: Multiwords and word sense disambiguation. In: *International Conference on Intelligent Text Processing and Computational Linguistics* Springer, 2005, S. 250–262

[CEM+17] Constant, Mathieu ; Eryiğit, Gülşen ; Monti, Johanna ; Van Der Plas, Lonneke ; Ramisch, Carlos ; Rosner, Michael ; Todirascu, Amalia: Multiword expression processing: A survey. In: *Computational Linguistics* 43 (2017), Nr. 4, S. 837–892

[FK11] Finlayson, Mark A. ; Kulkarni, Nidhi: Detecting multi-word expressions improves word sense disambiguation. In: *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World* Association for Computational Linguistics, 2011, S. 20–24

[Hey19] Hey, Tobias: INDIRECT: intent-driven requirements-to-code traceability. In: *Proceedings of the 41st International Conference on Software Engineering: Companion Proceedings* IEEE Press, 2019, S. 190–191

[McC09] McCarthy, Diana: Word sense disambiguation: An overview. In: *Language and Linguistics compass* 3 (2009), Nr. 2, S. 537–558

[Mih07] Mihalcea, Rada: Using wikipedia for automatic word sense disambiguation. In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 2007, S. 196–203

[PND07] Palmer, Martha ; Ng, Hwee T. ; Dang, Hoa T.: Evaluation of WSD systems. In: *Word Sense Disambiguation*. Springer, 2007, S. 75–106

[Wor] *WordNet Search - 3.1.* `http://wordnetweb.princeton.edu/perl/webwn`, . – Accessed: 2019-11-15