Relationsextraktion aus Fachsprache – ein automatischer Ansatz für die industrielle Qualitätsanalyse

Von Christian Hänig und Martin Schierle

Qualitätsanalysen in der Automobilindustrie basieren auf strukturierten Daten wie Schadensschlüsseln oder Bauteilnummern. Neben diesen Daten wird auch ein Freitext erfasst, welcher wichtige Informationen für die Qualitätsprozesse enthält. Dieser Beitrag präsentiert im Folgenden einen Workflow aus mehreren Sprachanalyse-Modulen zur Extraktion interessanter Begriffe (wie Bauteile und Symptome) und deren Relationen zueinander.

Schlagwörter

Relationsextraktion, unüberwachtes Parsen, Thesaurus

Einführung

Qualitätsanalysen in der Automobilindustrie verwenden normalerweise statistische Methoden und Algorithmen des maschinellen Lernens zur Ursachenanalyse, Frühwarnung sowie zur Trenderkennung. All diese Methoden setzen strukturierte Daten voraus, wie sie üblicherweise durch Fahrgestellnummern, Teilenummern oder Schadensschlüssel gegeben sind. Unstrukturierte Texte werden hierbei jedoch oft vernachlässigt oder ignoriert, meist aufgrund der hohen Analysekomplexität. Doch gerade Texte aus Internetforen, Callcenterprotokollen oder Werkstattberichten bieten ein nicht zu unterschätzendes Potential für die Qualitätsanalyse, da Kunden aus ihrer eigenen Sicht einen Fehler detailliert mit Randbedingungen und Erscheinungsformen beschreiben. Sobald die Texte von Angestellten in technische Codes übersetzt werden, gehen Informationen verloren oder werden schlicht falsch kodiert.

In Bezug auf die Qualitätsanalyse interessiert hauptsächlich, welche Komponenten von welchen Symptomen betroffen sind. Während sich Komponenten meist eindeutig über Terminologie erkennen lassen, ist die Detektion der sprachlich meist deutlich unschärfer oder gar metaphorisch formulierten Symptome deutlich schwieriger. Doch selbst wenn all diese Symptome eindeutig identifiziert wurden, ist die Zuordnung von Bauteilen zu Fehlern nicht einfach. Diese eindeutige Zuordnung von verschiedenen (benannten oder unbenannten Entitäten) zueinander wird als Relationsextraktion bezeichnet.

Das wissenschaftliche Interesse an Methoden der Informationsextraktion erwachte in den achtziger Jahren mit den Message Understanding Conferences (MUC, [1]). Für die ursprünglich militärisch orientierten Konferenzen mussten spezifische Informationen (wie z. B. Ort und Art eines terroristischen Anschlags) aus Texten extrahiert werden. Viele Ansätze zur Lösung nutzten Extraktionsmuster, welche auf den Text angewendet wurden und die Fakten von Interesse gezielt extrahierten. Eine gute Übersicht über musterbasierte Extraktionsverfahren findet sich in [2]. Extraktionsmuster können auch mit syntaktischen Informationen wie Wortarten angereichert oder unüberwacht bzw. semiüberwacht gelernt werden.

Viele der wissenschaftlichen Ansätze konzentrieren sich jedoch – sicher auch aufgrund der Vorgaben der Message Understanding Konferenzen – auf benannte Entitäten wie Personen-, Firmen- und Ortsnamen und Relationen zwischen diesen (z. B. „Welche Person wird von welcher Firma eingestellt?“). Zusätzlich wurden die meisten Algorithmen zur Anwendung auf Nachrichtentexte optimiert und evaluiert, wodurch eine Übertragung auf andere (Fach-)Sprachen erschwert wird. Dies ist insbesondere im Zeitalter des Web 2.0 ein Problem, da jedes Forum eine andere Domäne und jede Community eine andere Grammatik benutzt.

Die von uns vorgestellte Lösung zur Extraktion relevanter Information aus der Automobilfachsprache nutzt eine thesaurusbasierte Begriffserkennung zur Identifikation unbenannter Entitäten. Diese werden mit einer syntaktischen Analyse des Textes kombiniert, um so die gesuchten Relationen zwischen den Entitäten anhand einfacher Regeln auf den unüberwacht generierten Syntaxbäumen zu identifizieren.

Charakteristik automobiler Reparaturberichte

Die in dieser Arbeit verwendeten Reparaturberichte werden von Technikern in den Werkstätten direkt notiert. Dies geschieht parallel zum Gespräch mit dem Kunden und der Untersuchung des Autos, und führt daher zu einer ungewöhnlich hohen Zahl an Fehlern. So fließen häufig Rechtschreibfehler oder Tippfehler ein und es entstehen Segmentierungsfehler (fehlende oder zusätzliche Leerzeichen). Da die Texte unter großem Zeitdruck geschrieben werden, erhöht sich der Gebrauch von (u. U. mehrdeutigen) Abkürzungen, oder es werden zuweilen nur einzelne Buchstaben notiert. Durch das hochspezialisierte technische Umfeld weisen diese Texte eine hohe Dichte an Terminologie auf. Dabei treten Einwort- ebenso wie Mehrwortbenen-

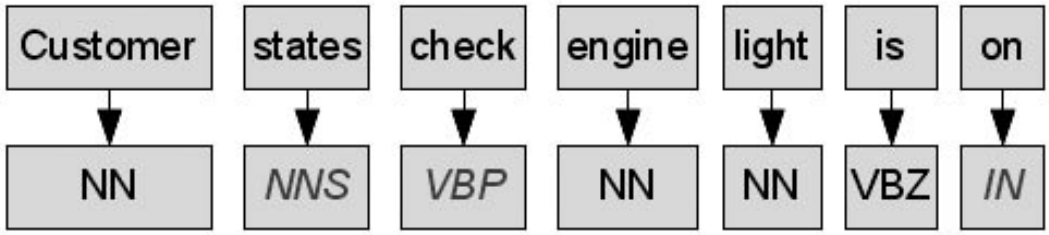


Abbildung 1: Beispiel für fehlerhaft annotierte Wortarten. Die Motorwarnleuchte (check engine light) wird durch den Stanford-Tagger nicht richtig erkannt und führt zu einer weiteren (states) Fehlklassifikation.

nungen auf, wodurch vor allem Wortart-Tagger, die auf anderen Korpora trainiert wurden, negativ beeinflusst werden (siehe Beispiel in Abbildung 1). Auch syntaktisch enthalten die Reparaturtexte eine eigene, spezielle Grammatik, was schon von Sprachregistern wie Computer Talk oder Baby Talk bekannt ist. Diese restringierte, stichwortähnliche Grammatik ist nur durch eigens dafür angelernte Parser zu verarbeiten, was angesichts mehrerer vorhandener Sprachen und Quellen nicht sehr zukunftssicher ist. Deshalb sollte ein System zur Verarbeitung dieser Daten möglichst auf unüberwachte Verfahren zurückgreifen, um das aufwendige Erstellen von Trainingsdaten zu vermeiden. Im Folgenden werden wir einen solchen Workflow zur Relationsextraktion vorstellen und auf die einzelnen Komponenten eingehen.

Ein System zur Informationsextraktion

Das System, welches zur Extraktion der relevanten Fakten verwendet wird, wurde als Verarbeitungskette von UIMA-Modulen konzipiert. UIMA (Unstructured Information Management Architecture) bezeichnet ein Framework, welches speziell zur Analyse und Verarbeitung unstrukturierter Daten (wie Freitext, Bild- oder Videodaten) von IBM entwickelt wurde. Mittlerweile wurde es von Apache in Apache Incubator integriert und von deren Community als Open-Source-Projekt weiterentwickelt.

Vorteile von UIMA bei der Analyse unstrukturierter Daten sind vor allem die hohe Modularität des Systems, in welchem jeder Baustein durch eine zusätzliche Deskriptor-Datei beschrieben wird. Hier können sowohl (sprachabhängige) Parameter des Moduls definiert, als auch Vorbedingungen, Nachbedingungen und verwendete Ressourcen spezifiziert werden. Durch die Angaben lassen sich Sprachanalysemodule schnell und unkompliziert zu sprach- und domänenunabhängigen Workflows kombinieren. Sowohl Ressourcenverwaltung, Parameterkonfigurationen als auch Verteilung auf mehrere Rechner kann dann von UIMA übernommen werden.

Ein weiterer großer Vorteil durch die Verwendung von UIMA stellt die Vielzahl fertig implementierter Module dar, die über entsprechende Repositories durch Universitäten und andere Institutionen zur Verfügung gestellt werden.

Textaufbereitung

Die ersten Module des Systems dienen der Textaufbereitung und führen sowohl Sprachenidentifikation, Wortsegmentierung, als auch Rechtschreibkorrektur und Abkürzungsersetzung durch. Die Wortsegmentierung geschieht über eine Liste von manuell erstellten regulären Ausdrücken und identifiziert besondere Angaben wie Zahlen, Kilometer- und Preisangaben oder auch interne Fehlercodes. Diese werden später für das semantische Taggen herangezogen.

Für die Rechtschreibkorrektur wurde der bekannte Korrekturalgorithmus von ASpell weiterentwickelt. Neben Frequenzinformationen im Wörterbuch wurde der Algorithmus um Kontextinformationen durch Nachbarschaftskookkurrenzen ergänzt, sodass Korrekturen kontextabhängig durchgeführt werden können. Eindeutige Abkürzungen werden über einfache Abkürzungslisten ersetzt, während mehrdeutige Abkürzungen – ähnlich wie bei der Rechtschreibkorrektur – über den Kontext aufgelöst werden.

Thesaurusbasierte Begriffserkennung

Im Gegensatz zu anderen Arbeiten zur Relationsextraktion geht es bei der Analyse von Reparaturberichten nicht um benannte Entitäten wie Personennamen, sondern um Symptome jeder Art. Diese können sich nicht nur in Terminologien widerspiegeln, sondern auch in allgemeinsprachlichen Ausdrücken oder sogar Phrasen. Für spätere Analyseschritte (Auswertungen usw.) ist es wichtig, Benennungen ein und desselben Begriffs zusammenzufassen und mehrdeutige Benennungen aufzulösen. Des Weiteren ist die Hinterlegung von Hyponym- bzw. Hyperonymbeziehungen und Meronymbeziehungen essenziell.

Diese Anforderungen wären bei einem Thesaurus wie z. B. WordNet erfüllt, doch die Mehrsprachigkeit der Texte erfordert neben der Definition von Synonymie, Meronymie und Hyperonymie auch die Definition multilingualer Beziehungen. Zusätzlich muss die Datenstruktur auch zur Identifikation der Benennungen im Text beitragen, was bei WordNet (oder dem mehrsprachigen EuroWordNet) nicht ausreichend gegeben ist.

Um all diese Anforderungen zu erfüllen, wird in dem Extraktionssystem ein speziell entwickelter mehrsprachiger Thesaurus verwendet, welcher sowohl Terminologie als auch für die Anwendung relevante Alltagssprache enthält.

Jede abstrakte Bedeutung wird als eigener Begriff modelliert und kann mehrere Synonyme in mehreren Sprachen haben. Da Benennungen nur in bestimmten Kontexten als synonym (bzw. direkte Übersetzung) zu erachten sind, kann jedes Token einer Begrifflichkeit durch Wortarten und Kontextschlüsselwörter

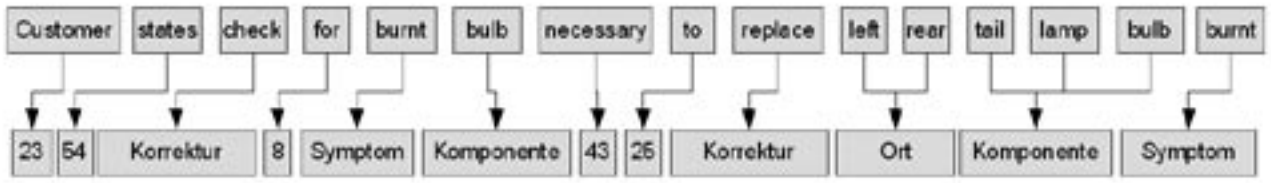


Abbildung 2: Reparaturbericht mit semantischen Tags

angereichert werden. Dies ermöglicht es, ein Wort in Abhängigkeit von seinem Kontext auf einen abstrakten Begriff abzubilden und realisiert so eine direkte Zuordnung zwischen den verschiedenen Sprachen. Flexionsformen müssen nicht umfassend in den Thesaurus eingepflegt werden, sondern können über reguläre Ausdrücke abgedeckt werden. Dies vermeidet die Verwendung einer fehlerträchtigen und sprachabhängigen Grundformreduktion. Zusätzlich werden alle Kontexte in einer polyhierarchischen Struktur organisiert, die Meronymie- und Hyperonymiebeziehungen unterstützt.

Die Pflege des Thesaurus wird durch eine automatische Synonymerweiterung vereinfacht. Falls Einzelwörter einer Mehrwortbenennung bereits mit Synonymen eingepflegt wurden, werden alle möglichen Mehrwortbenennungen mit diesen Synonymen gebildet und ergänzt. So wird *brake light* automatisch aus *brake lamp* inferiert, wenn *light* und *lamp* als Synonyme hinterlegt wurden. Obwohl dies die Ableitung unwahrscheinlicher Ausdrücke ermöglicht, entstehen erfahrungsgemäß trotzdem keine falschen Einträge.

Dieser Thesaurus bildet die Grundlage für die Begriffserkennung. Nach Anwendung eines Wortart-Taggers und Vorindizierung aller Benennungen in einem Präfixbaum (Trie-Struktur) zum effizienten Suchen und Vergleichen können alle relevanten Begriffe der Domäne einheitlich und eindeutig identifiziert werden [3].

Semantisches Taggen

Für syntaktische Parser werden normalerweise linguistische Kategorien in Form von Wortarten herangezogen, da diese Informationen über die Funktionen der Worte im Satz enthalten. Für eine qualitativ hochwertige Relationsextraktion sind strukturelle Analysen der Sprache essenziell, jedoch können die genutzten Kategorien noch um semantisches Wissen erweitert werden. Dafür werden Benennungen, welche in dem Thesaurus enthalten sind, durch ein semantisches Tag ersetzt, das die entsprechende Kategorie kennzeichnet (z. B. wird *Bremspedal* durch das Tag *Komponente* ersetzt). Danach findet die Ersetzung von besonderen Angaben wie Zahlen, Kilometer- und Preisangaben, Datums- und Uhrzeitangaben und internen Fehlercodes durch geeignete Platzhalter statt.

Die Verwendung semantischer Tags hat zwei positive Effekte. Zum einen werden seltene Benennungen statistisch besser repräsentiert, da durch die Ersetzung mit einer übergeordneten Kategorie mehr Belegstellen zur Analyse herangezogen werden können. Infolge dessen werden rund 83% aller Wörter durch entsprechende Tags ersetzt, wodurch eine Abstraktion von der zugrunde liegenden Sprache stattfindet. Diese Abstraktion vereinfacht die nachfolgende syntaktische Analyse erheblich.

Für gute Strukturanalysen ist vor allem eine hohe Qualität der Wortart-Tags notwenig. Da der Sprachgebrauch in verschiedenen Texttypen wie Reparaturberichten, Forenbeiträgen oder Zeitungstexten sehr voneinander abweicht, sollte je Sprache und Texttyp ein Modell berechnet werden. Nur so kann eine Fehlerfortpflanzung vom Wortart-Tagging zum Parsen minimiert werden. Leider stehen für die verschiedenen Kombinationen von Texttyp und Spra-

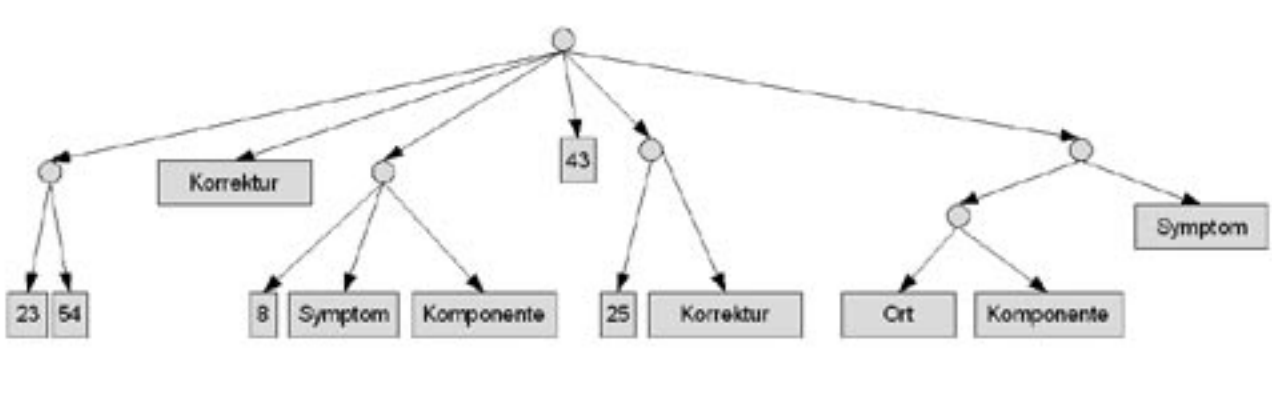


Abbildung 3: Syntaxbaum der semantischen Tags des Reparaturberichtes

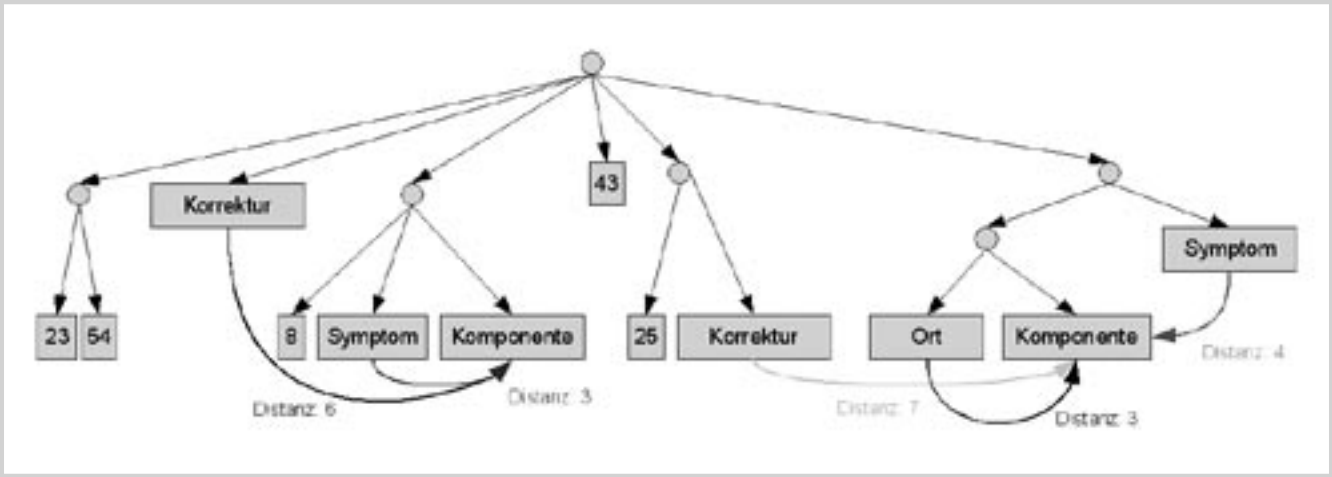


Abbildung 4: Syntaxbaum der semantischen Tags des Reparaturberichtes

che keine geeignet annotierten Korpora zur Verfügung, wodurch das Trainieren der meisten Wortart-Tagger unmöglich wird, wenn nicht sehr viel Zeit und Geld in das Annotieren neuer Textressourcen investiert werden soll. Unüberwachte Verfahren sind daher besser geeignet, da sie als Trainingsmenge lediglich eine große Menge an Text benötigen. Die Wörter werden durch eine Clusteranalyse – basierend auf Kontextähnlichkeiten – in verschiedene Wortklassen gruppiert, welche mit Wortarten vergleichbar sind [4]. Da bei diesem unüberwachten Tagger kein Wissen über Wortarten bzw. die Zugehörigkeit der resultierenden Cluster zu Wortarten vorliegt, werden die extrahierten Wortgruppierungen durchnummeriert. Ein Beispiel für einen mit semantischen Tags angereicherten Text ist in Abbildung 2 zu sehen (die Cluster, die die vorher ersetzten semantischen Tags enthalten, wurden zur besseren Lesbarkeit durch ihr Label repräsentiert).

Syntaktisches Parsen

Durch die Anwendung einer manuell erstellten Grammatik (bzw. annotierten Korpora) lassen sich Texte mit Syntaxbäumen anreichern, welche eine wichtige Grundlage der Relationsextraktion darstellen. Durch das Markieren von interessanten Benennungen im Syntaxbaum können nicht nur Aussagen über Zuordnungen getroffen werden, sondern auch Relationen beliebiger Stelligkeit (wie z. B. [Komponente, Symptom] oder [Komponente, Symptom, Bedingung]) betrachtet werden. Wie beim Wortart-Tagging fehlen in der Automobildomäne jedoch geeignete Trainingsdaten in Form von mit syntaktischen Informationen angereicherten Korpora, sodass auch hier auf ein vollautomatisches Verfahren gesetzt wird. Bei Daimler fiel die Wahl auf unsuParse ([5]), da dieser Algorithmus nicht nur die genauesten Ergebnisse liefert, sondern auch eine schnelle Laufzeit bietet. Dies ist angesichts der großen Datenmengen, welche tagtäglich anfallen, von besonderer Bedeutung. Aufbauend auf den semantischen Tags werden durch Nachbarschaftskookkurrenzen statistische Maßzahlen berechnet, die einen Trennwert zwischen den angrenzenden Tags darstellen. Durch diese Werte wird erkannt, welche Tags im Textfluss syntaktisch zu einer Phrase gehören und welche eine weniger starke Bindung zueinander besitzen. Diese Informationen werden iterativ genutzt, um automatisch einen Syntaxbaum des Satzes zu erstellen. Da die interessanten Entitäten unserer Domäne oft in syntaktischer Relation zueinander stehen (Substantiv und Modifikator etc.), ist ihre Distanz im Syntaxbaum gering, was direkt zur Extraktion der Relationen genutzt werden kann. Der entsprechende Syntaxbaum für das obige Beispiel ist in Abbildung 3 zu sehen.

Regelbasierte Relationsextraktion

Der nächste und letzte Schritt im vorgestellten Workflow stellt die eigentliche Extraktion von Wissen dar. Da die Entitäten von Interesse normalerweise in syntaktischer Relation stehen, ist die eigentliche Zuordnung nach Erstellung des Syntaxbaumes einfach und kann über wenige manuell erstelle Regeln abgehandelt werden. Für jede zu untersuchende Relation wird eine Regel definiert. Diese besteht aus den beteiligten Begriffskate-

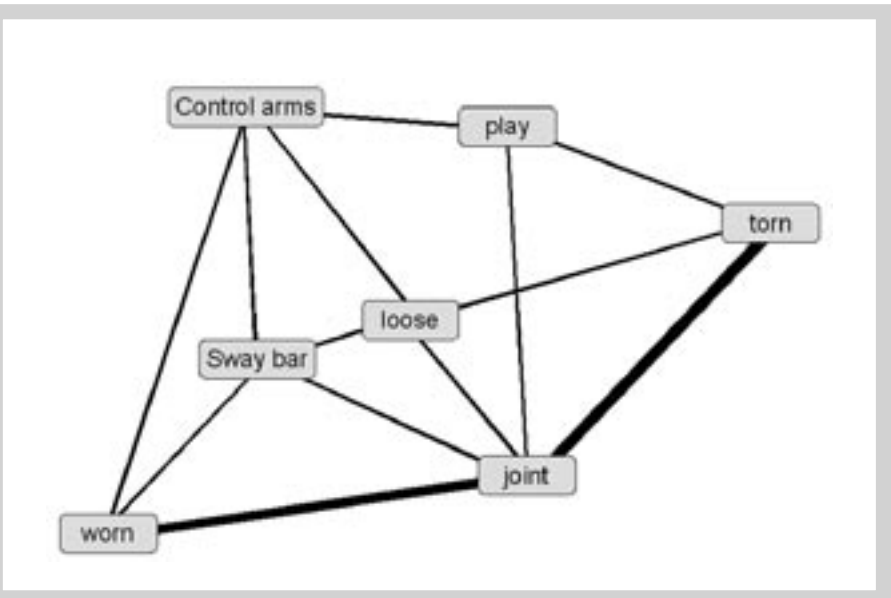
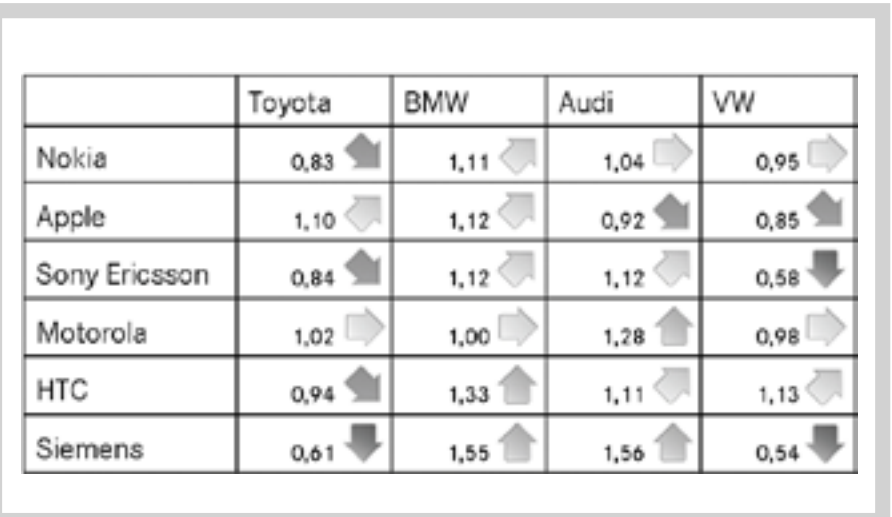


Abbildung 5: Symptomgraph aus extrahierten Relationen zur Ursachenanalyse



	Toyota	BMW	Audi	VW
Nokia	0,83	1,11	1,04	0,95
Apple	1,10	1,12	0,92	0,85
Sony Ericsson	0,84	1,12	1,12	0,58
Motorola	1,02	1,00	1,28	0,98
HTC	0,94	1,33	1,11	1,13
Siemens	0,61	1,55	1,56	0,54

Abbildung 6: Wettbewerbsvergleich – Stimmungsanalyse bezüglich Handyanbindungen im Auto

gorien sowie drei Parametern zur Extraktion aus dem Syntaxbaum. Diese Parameter werden heuristisch festgelegt und beschreiben die bevorzugte Richtung der Relation und den maximalen Abstand (Summe aus Abstand zwischen den Worten und Pfadlänge im Syntaxbaum) zwischen den Entitäten. Zusätzlich legt ein Marker fest, ob diese Relation auch ein Teil einer komplexeren Relation sein kann. Während ein erster Schritt also nur binäre Relationen erkennt, können in weiteren Analyseschritten diese zu n-ären Relationen erweitert werden. Die Anwendung der Regeln auf den Syntaxbaum entspricht der Extraktion der gesuchten Relationen. Ein Beispiel stellt die Abbildung 4 dar, in der die erkannten Relationen mit der ihnen zugeordneten Distanz hervorgehoben sind.

Zusammenfassung und Ausblick

Zur Evaluation und qualitativen Analyse des vorgestellten Workflows zur Relationsextraktion wurden 100 Texte zufällig ausgewählt. Diese wurden manuell analysiert, die enthaltenen Relationen annotiert und mit den Resultaten des automatischen Systems verglichen. Dabei erzielte der vorgestellte Algorithmus eine Vollständigkeit von 80%, wobei über 87% der extrahierten Relationen korrekt waren. Werden Fehler ignoriert, die aus Fehleinträgen (bzw. fehlenden Einträgen) im Thesaurus resultieren, erreicht der Algorithmus sogar eine Vollständigkeit bzw. Korrektheit von 95%. Dies zeigt zum einen die sehr gute Performanz des vorgestellten Systems und auf der anderen Seite den positiven Einfluss gut gepflegter fachspezifischer Thesauri. Da diese durch Wartung erweiterbar sind, kann das System durch den Anwender selbst für neu entstehende Themen schnell und leicht angepasst werden.

Dazu trägt die Integration von unüberwachten und vollautomatischen Verfahren für die Sprachverarbeitung bei. Gerade in hochspezialisierten Domänen wie z. B. der Automobilbranche liefert der vorgestellte Workflow einen Beitrag zur effizienten und kostengünstigen Informationsgewinnung, da keine Trainingsdaten für die einzelnen Module zeit- und kostenaufwendig erstellt werden müssen. Lediglich das schon vorhandene Fachvokabular für den entsprechenden Anwendungsbereich fließt ein und kann bei Bedarf erweitert werden.

Durch vollautomatische Textverarbeitung und Wissensextraktion können neue und schnell wachsende Quellen (wie z. B. das Web 2.0) erschlossen werden und zur Qualitätsanalyse und somit auch -verbesserung herangezogen werden. Während die extrahierten Relationen einer größeren Textkollektion zur Ursachenanalyse in Form von Graphen herangezogen werden können (siehe Abbildung 5), liefern Stimmungsanalysen der Kunden wertvolle Informationen bezüglich Zufriedenheit und Vergleich mit Wettbewerbern (siehe Abbildung 6).

In der Zukunft werden wir untersuchen, inwiefern Fachvokabular automatisch gelernt und in eine Wissensstruktur wie den vorgestellten Thesaurus überführt werden kann. Dies soll den manuellen Aufwand weiter verringern und somit eine schnelle Adaption des Workflows auf andere Domänen und Themen erleichtern.

Literatur

[1] Grishman, R.; Sundheim, Grishman, R. / Sundheim, B. (1996): Message understanding conference - 6: A brief history. In: Proceedings of the International Conference on Computational Linguistics.

[2] Muslea, I. (1999): Extraction patterns for information extraction tasks: A survey. In: In AAAI-99 Workshop on Machine Learning for Information Extraction.

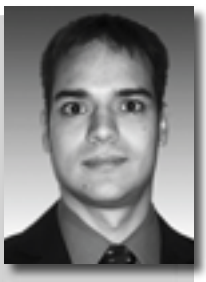
[3] Schierle, M. / Trabold, D. (2008): Multilingual knowledge based concept recognition in textual data. In: Proceedings of the 32nd Annual Conference of the GfKl.

[4] Biemann, C. (2006): Part-of-speech tagging employing efficient graph clustering. In: Proceedings of the COLING/ACL-06 Student Research Workshop.

[5] Hänig, C. / Bordag, S. / Quasthoff, U. (2008): Unsuparse: Unsupervised parsing with unsupervised part of speech tagging. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).

Christian Hänig absolvierte sein Informatikstudium an der Universität Leipzig. Seitdem ist er in der Forschung und Entwicklung der Daimler AG in der Qualitätsanalyse als Doktorand tätig. Der Schwerpunkt seiner Forschungsarbeit liegt im Bereich der syntaktischen Analyse natürlicher Sprache und Relationsextraktion.

Kontaktadresse
Daimler AG
christian.haenig@daimler.com
www.daimler.com



Martin Schierle machte seinen Abschluss als Diplom-Informatiker an der Universität Ulm. Seitdem ist er als Doktorand in der Forschung und Entwicklung der Daimler AG tätig. Er beschäftigt sich mit der Qualitätsanalyse auf unstrukturierten Daten. Sein Schwerpunkt liegt dabei auf Systemmodellierung und Relationsextraktion.

Kontaktadresse
Daimler AG
martin.schierle@daimler.com
www.daimler.com