Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Ausarbeitung Master Seminar 2 SS 2014

Ivan Demin

Text Mining for Second Screen

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Ivan Demin

Text Mining for Second Screen

Ausarbeitung Master Seminar 2
SS 2014 eingereicht im Rahmen der Master Seminar 2

im Studiengang Master Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Gutachter: Prof. Dr. Bettina Buth
Gutachter: Prof. Dr. Kai von Luck

Betreuer: Prof. Dr. Kai von Luck

Eingereicht am: 30. August 2014

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1. Einleitung

Second Screen ist seit einigen Jahren ein Bestandteil der aktuellen Mediendebatten. Im Rahmen der Arbeit [Dem14] wurde ein erster Überblick über die Thematik gegeben. Second Screen steht in erster Linie für die parallele Nutzung eines internetfähigen Geräts wie Smartphone, Tablet und Desktop Rechner mit Bezug zum laufenden Fernsehprogramm. Dabei teilt es sich in die Komponenten First Screen, welche zum Konsumieren von Medieninhalten gedacht ist, sowie Second Screen die Kontext zu dem Geschehen liefert, auf. In zweiter Linie bietet Second Screen die Möglichkeit sich durch Soziale Medien wie Twitter und Facebook mit weiteren Medienkonsumenten über das Gesehene auszutauschen.

Second Screen Anwendungen werden bei den deutschen Nutzern immer beliebter. Nach einer repräsentativen Onlinestudie von ARD und ZDF aus dem Jahr 2012 haben 28 % der deutschen Fernsehzuschauer angegeben, sich begleitend zur Fernsehsendung im Internet informiert zu haben [EF12]. In der ein Jahr darauf folgenden Untersuchung waren es schon 33 % der deutschen Fernsehzuschauer [EF13]. Dagegen lag die Second Screen Nutzung in den USA konstant bei über 80 % im Jahr 2012 und 2013 [Nie12, Nie14]

Nicht jeder Anwendungsfall macht im Second Screen Bereich Sinn. Mehrere Studien belegen, dass die Menschen nur bedingt multitaskingfähig sind [GM04, STB09]. Demnach braucht das menschliche Gehirn eine gewisse Verzögerungszeit, um auf eine andere Aufgabe umzuschalten [TABM03]. Damit muss bei dem Nutzer eine ausreichend lange Aufmerksamkeitsspanne gegeben sein. Kurze Beiträge wie Nachrichtenblöcke eignen sich dafür nicht, wogegen längere Videos wie Dokumentationen dem Benutzer nach einem Wechsel einen ausreichenden Zeitraum bieten sich zurück ins Geschehen einzufinden.

Demnach ist ein Dokumentationsanwendungsfall möglich, wonach auf dem First Screen der eigentliche Beitrag läuft und der Second Screen dabei den Kontext zu dem Geschehen zur Verfügung stellt. Der Kontext teilt sich in geografische, kritische und ergänzende Informationen auf. Ziel der Ausarbeitung ist es eine Grundlage für die Realisierung des ***I**nformation **f**or **S**econd **S**creen-Systems* (IfSS) für die Umsetzung eines Dokumentations-Anwendungsfalls zu schaffen. Hierfür muss eine Kontextkomponente realisiert werden, welche aus den Dokumentationsbeschreibungen Anfragen an Informationsquellen generieren kann. Demnach muss eine Informationsquelle gefunden werden, welche ein breites Themenfeld abdeckt, aus verifizierten Informationen besteht, semistrukturiert ist und einen einfachen Zugang auf die Daten bietet. Eine Gesamtübersicht der möglichen zu realisierenden Komponenten des IfSS-Systems kann aus A.1 entnommen werden.

2. Relevante Ansätze

Der folgende Abschnitt beschäftigt sich mit drei Ansätzen zur Erkennung von Schlagwörtern aus Texten als Zusatzinformationen. Die Herangehensweise bei dem Trainieren und Evaluieren der Klassifikatoren wird erläutert und Rückschlüsse für die Verwendung bei der Konzipierung der Kontext-Komponente werden gezogen.

2.1. Learning to link with wikipedia

Das folgende Paper beschreibt ein Verfahren aus algorithmisch-anwendungsorientierter Perspektive zur automatischen Kreuzreferenzierung von Dokumenten mit Wikipedia-Artikeln nach Milne et al. [MW08]. Es ist ein Weg mithilfe maschinellen Lernens Zusatzinformationen anhand von Links zu Wikipedia-Artikeln bereitzustellen. Schon einige Arbeiten implementierten den Ansatz für die Generierung von strukturierten Anfragen an Informationssysteme [JM09, CGL+13].

Um die Kreuzreferenzierung zu ermöglichen, wollten die Autoren das Wikifizierungs-Verfahren[1] von Wikipedia erlernen. Dieses Verfahren wird dazu genutzt, um möglichen Verständnisschwierigkeiten der Nutzer vorzubeugen. Demnach werden die Wikipedia-Artikel miteinander durch eine Verlinkung verbunden, welche semantisch in Verbindung zueinander stehen. Dazu wurden aus zwei Millionen Artikeln 700 Artikel rausgesucht, die mindestens eine Verlinkung zu 50 Artikeln aufweisen. Die gesammelten Artikel wurden als Trainingsdaten und Testdaten verwendet. Als Klassifikatoren wurden mehrere Klassifikationsverfahren, darunter der auf Häufigkeit basierte Naiver Bayes, die lineare Support Vektor Maschine und der Entscheidungsbaum C4.5 verwendet.

Das Verfahren splittet sich in Link-Begriffsklärungsphase und Link-Ermittlungsphase auf. In der ersten Phase wird die Mehrdeutigkeit der möglichen Linkkandidaten aufgelöst. Dabei können mögliche Linkkandidaten eine doppeldeutige Bedeutung haben. So kann zum Beispiel das Wort Note als Banknote oder Musiknote verstanden werden. Zu jedem der beiden Bedeutungen wird ein Wikipedia-Artikel angeboten.

Link-Begriffsklärungsphase

Um die Sinnhaftigkeit der einzelnen Bergriffe festzustellen, werden die beiden Features *Commonness* und *Relatedness* verwendet. Die *Commonness* beschreibt die Häufigkeit, wie oft auf den Wikipedia-Artikel von Wikipedia im Vergleich zu den anderen möglichen Bedeutungen der Wörter verwiesen wird. *Relatedness* dagegen beschreibt die Verbundenheit der Linkkan-

[1] http://de.wikipedia.org/wiki/Wikipedia:Wikifizieren

didaten mit dem Artikel, dazu werden die eingehenden und ausgehenden Links der beiden Artikel untersucht. Desto höher der gemessene Wert ist, desto höher ist die semantische Gemeinsamkeit des Linkkandidaten zu dem untersuchten Artikel.

Link-Ermittlungsphase

Die Link-Ermittlungsphase geht der Frage nach, welche Terme verlinkt werden sollten. Dazu wurde ein fünfteiliges Feature-Set bestehend aus Linkswahrscheinlichkeit, Verbundenheit, Konfidenz der Begriffsklärung, Allgemeingültigkeit und Lokation definiert.

Die Linkswahrscheinlichkeit beschreibt die Gesamtwahrscheinlichkeit für einen Link. Hierbei haben Linkkandidaten, welche aus mehreren Wörtern bestehen eine höhere Linkswahrscheinlichkeit zugesagt. Damit würde die Wortkombination "theoretische Informatik"besser bewertet, als nur das Wort Informatik. Die Verbundenheit beschreibt die Beziehung der Linkkandidaten zu allen ausgewählten Linkkandidaten. Demnach werden die eingehenden und ausgehenden Links untersucht. Desto größer die Gemeinsamkeit der Verlinkungen des Linkkandidaten mit dem untersuchten Artikel haben, desto höher ist seine Verbundenheit zu dem Artikel. Die Konfidenz der Begriffsklärung beschreibt den kombinierten Durchschnitt und maximalen Wert dafür, dass der Link in Beziehung zum Kontext steht. Das nachfolgende Feature beschreibt die Allgemeingültigkeit des Linkkandidaten. Es definiert den minimalen Abstand vom Thema des untersuchten Wikipedia-Artikels zu dem Kandidaten als aufgespannte Baumstruktur. Das letzte Feature der Lokation beschreibt die Position des Linkkandidaten im Artikel. Demnach haben Verlinkungen in der Einleitung und am Ende des Textes einen höheren Wert für die Verständlichkeit [DL95].

Untersuchung

Da der Klassifikator C4.5 bessere Zwischenergebnissen als Naive Bayes und SVN lieferte, wurde er für die Abschlussuntersuchung verwendet. Die endgültige Untersuchung des trainierten Klassifikatoren erfolgte anhand von 50 ausgewählten Nachrichtenartikeln. Hierbei mussten 88 Probanden die Korrektheit der durch die Verwendung des trainierten Klassifikators generierten Links bestimmen. Die Tabelle 2.1 zeigt die Ergebnisse der durchgeführten Untersuchung. Die generierten Verlinkungen waren zu 76,4 % als korrekt und zu 23,6 % falsch klassifiziert worden.

Tabelle 2.1.: Performance vom Algorithmus für die Linkerkennung [MW08]

correct	76.4
incorrect (wrong destination)	0.9
incorrect (irrelevant and/or unhelpful)	19.8
incorrect (unknown reason)	2.9

Fazit

Das vorgestellte Verfahren stellt einen interessanten Ansatz zum Erlernen des Wikifizierung-Verfahrens dar. Für das umzusetzende IfSS-System ist die Beseitigung der Mehrdeutigkeit aus den Texten unabdingbar. Hierfür stellen die verwendeten Merkmale *Commonness* und *Relatedness* eine solide Basis dar. Die unterschiedlichen Bedeutungen mit dem dazugehörigen Wikipedia-Artikel, sowie die Gesamtzahl der Verlinkungen, kann durch die MediaWiki-Api [2] geprüft werden. Bei jeder einzelnen Bedeutung aus der Liste muss eine Anfrage an Wikipedia gesendet werden, was einen hohen Kommunikationsaufwand nach sich zieht. Die Möglichkeit der Zwischenspeicherung von Ergebnisse muss im späteren Verlauf des Projektes betrachtet werden.

Des Weiteren stellt das aufgestellte Feature-Set eine gute Grundlage zur Identifizierung von Linkkandidaten in Texten bereit. Damit können alle Merkmale später bei der Entwicklung des Klassifikators für die Kontext-Komponente verwendet werden. Hinzu kommt, das Wikipedia durch die MediaWiki-Api eine einfache Möglichkeit bietet, alle Wikipedia-Artikel gepackt herunterzuladen und somit an verifizierte Trainings und Testdaten zu gelangen [3].

Dennoch müssen die erzielten Ergebnisse der Evaluierung mit Vorsicht betrachtet werden. Da nur 88 Probanden teilgenommen haben, ist die Untersuchung nicht repräsentativ. Das Ergebnis ist mit einer Erfolgsrate von 76 % nicht zufriedenstellend, da dem Nutzer noch zu viele falsche Links angeboten werden. Das Verfahren kann anhand der Erweiterung des Feature-Sets durch zum Beispiel der Hinzunahme des Traffic-Faktors im Feature-Set verbessert werden.

2.2. Feeding the second screen: semantic linking based on subtitles

Das nachfolgende Paper beschreibt die Generierung von Wikipedia-Links aus Untertiteln von Talkshows als Zusatzinformationen für den Second Screen nach Odijk et al. [OMR13]. Die Daten werden dabei als Kontextgraph [FR09] modelliert. Aufbauend auf den Erkenntnissen aus [MWR12] nachdem die Autoren zu Twitter-Post Semantik anhand von Kreuzreferenzierung von Schlagwörtern erzeugt haben. Übernahmen sie zu größten Teilen das verwendete Feature-Set. Des Weiteren verwenden sie aufgrund der besseren Ergebnisse gegenüber Support Vektor Maschinen das Klassifikationsverfahren der Entscheidungsbäume *Random Forest* [BRE01].

Der entwickelte Algorithmus arbeitet in drei Schritten: es werden zum Kontext passende Linkkandidaten gefunden, ihnen wird ein Wert zugewiesen und die Reihenfolge im Kontext-

[2] http://en.wikipedia.org/w/api.php
[3] http://dumps.wikimedia.org/backup-index.html

graphen gegebenenfalls angepasst. Hierfür werden die Untertitel in einzelne Teile t_i aufgeteilt. Jedem gefundenen Linkkandidaten L wird ein Fixpunkt a zu einem bestimmten Wikipedia-Artikel w zugeordnet.

Im zweiten Schritt wird jedem der Linkkandidaten ein Wert zugeordnet. Zu diesem Zweck wird die Häufigkeit, wie oft auf einen bestimmten Wikipedia-Artikel verwiesen wird, gemessen:

$$COMMONNESS(a, w) = \frac{|L_{a,w}|}{\sum_{w' \in W} |L_{a,w'}|'} \tag{2.1}$$

Der Grundgedanke davon ist, dass Linkkandidaten mit Fixpunkten, welche zum selben Wikipedia-Artikel verweisen, passendere Repräsentationen sind als Linkkandidaten die oft auf andere Ziele verweisen.

Umsortierung

Für die Umsortierung der Linkkandidaten in dem zu modellierenden Kontextgraphen wird ein 26-teiliges Feature-Set verwendet. Unterteilt wird es in vier Hauptteile bestehend aus: Fixpunkt-, Ziel-, Kombination aus Fixpunkt- und Ziel-, sowie Kontext-Feature.

Die Fixpunkte-Features beziehen sich auf den extrahierten Linkkandidaten in einem Textteil. Dazu wird unter anderem die Anzahl der Wikipedia Artikel mit dem gleichen Titel wie der Fixpunktname mit betrachtet. Des Weiteren wird die Eintrittswahrscheinlichkeit, dass der Name vom Fixpunkt bei Wikipedia verwendet wird, einbezogen.

Dagegen beziehen sich die Ziel-Features auf den eigentlichen gefundenen Wikipedia-Artikel. Hierfür wird die Anzahl der Verweise von Wikipedia selbst, sowie die Gesamtanzahl der Verlinkungen auf den Wikipedia-Artikel in der Berechnung verwendet. Außerdem wird Besucherzahl der Seite in den zeitlichen Perioden: Woche, Monat und Jahr mit betrachtet.

Im nächsten Hauptteil wird die Kombination aus Fixpunkt und Zielpunkt-Feature betrachtet. Hierfür wird untersucht ob der Titel vom gefundenen Wikipedia-Artikel im Fixpunkt enthalten ist, sowie die Umkehrung davon.

Der dritte Hauptteil betrachtet den Kontext in dem der Kontextgraph mit in die Bewertung hinzugezogen wird. Dazu wird die Anzahl der Kanten, sowie die Zentralität im Kontextgraphen zu dem Wikipedia-Artikel ermittelt. Als letztes wird die Metrik PageRank, welche vergeben wird um die Bedeutung einer Seite im Web zu bestimmen [PBMW98], hinzugezogen.

Kontextmodellierung

Die Daten wurden von den Autoren als Kontextgraph modelliert. Die Abbildung 2.1 zeigt einen Ausschnitt aus dem Kontextgraphen, bestehend aus drei extrahierten Textteilen (t_1, t_2, t_3) und

zwei verschiedenen Fixpunkten ((t_2, a) und (t_3, a')), welche zu einem bestimmten Wikipedia-Artikel w verweisen.

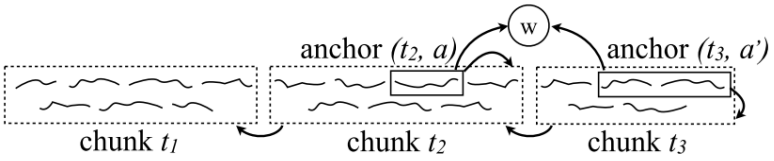

Abbildung 2.1.: Ausschnitt aus dem Kontextgraphen [OMR13]

Das Verfahren zur Erstellung vom Kontext-Graphen erfolgt in drei Schritten. Im ersten Schritt werden für jeden Textteil (*chunks*) mögliche Links generiert. Danach wird anhand der platzierten Grenze entschieden ob ein Link hinzugefügt wird oder nicht. Im letzten Schritt wird geprüft ob sich die Anzahl der Verlinkungen und damit das Kantengewicht zu einem bestimmten Wikipedia-Artikel (*w*) als Knoten ändert. So wird die Position des Knotens im Kontext-Graphen geändert.

Untersuchung

Für die Untersuchung wurden aus sechs Talkshow-Episoden Untertitel mit rund 36 tausend Wörtern als Test-Set verwendet. Für die extrahierten Textteile aus dem Test-Set wurden mit Hilfe des trainierten Klassifikators Links generiert. Anschließend wurde ein Goldstandard [4] etabliert, nachdem ein Annotationsspezialist manuell Links aus dem Test-Set generiert hat. Die Ergebnisse der eingesetzten Verfahren wurden danach mit den Links aus dem Goldstandard verglichen. Eine Auflistung der erreichten Werte kann aus der Tabelle 2.2 entnommen werden.

Tabelle 2.2.: Zusammenfassung der semantischen Verlinkungs-Ergebnisse aus der Untersuchung [OMR13]

	Average classification time per chunk (in ms)	R-Prec	MAP
1. Baseline retrieval model	54	0.5753	0.6235
2. Learning to rerank approach	99	0.7177	0.7884
Learning to rerank (L2R) + one context graph feature			
3. L2R+*DEGREE*	104	0.7375	0.8252
4. L2R+*DEGREECENTRALITY*	108	0.7454	0.8219
5. L2R+*PAGERANK*	119	0.7380	0.8187
Learning to rerank (L2R) + three context graph features			
6. L2R+*DEGREE*+*PAGERANK* +*DEGREECENTRALITY*	120	0.7341	0.8204

[4] Goldstandard: Ist ein Ausdruck aus der wissenschaftlichen und medizinischen Umgangssprache. Beschreibt ein Verfahren, welches unter konkurrierenden Verfahren in allgemeiner Auffassung als das Beste gilt.

Der Basisansatz hat mit einer durchschnittlichen Genauigkeit von 62 % richtige Ergebnisse geliefert. Durch das Hinzuziehen der aufgestellten Merkmale konnte die durchschnittliche Genauigkeit um rund 17 % gesteigert werden. Das Hinzuziehen der Kontextmerkmale erbrachte eine durchschnittliche Verbesserung von 21 %, wogegen die Kombination aller Kontextmerkmale keine weitere Verbesserung der Genauigkeit gebracht hatte.

Fazit

Das vorgestellte Verfahren stellt eine interessante Erweiterung des Ansatzes aus [MW08] dar. Dabei ist die Modellierung vom Kontext als Graphen nachahmenswert. Hierbei ergibt sich die Möglichkeit, durch das Kantengewicht und Position im Graphen einen Bezug zu dem Geschehen aufzubauen. Der untersuchte trainierte Klassifikator lieferte mit rund 82 % einen guten Wert. Dennoch müssen die Ergebnisse der Untersuchung mit Vorsicht betrachtet werden, da die Etablierung des Goldstandars durch den Annotationsspezialisten im Paper nicht nachvollziehbar ist. Das verwendete Feature-Set mit der Einbezugnahme der Besucherzahlen einer Unterseite entspricht dem eigentlichen Grundgedanken von Second Screen, nachdem die Aufrufe für eine Medienbeschreibungen während der Ausstrahlung einer Serie oder eines Films im Fernsehen steigen. Die Verwendung der im Internet etablierten Metrik PageRank als Feature eröffnet neue Möglichkeiten der Bewertung einer Information. So können für das aufzustellende Feature-Set der Kontext-Komponente weitere etablierte Internetmetriken wie der Alexa Rank[5] zur Bewertung von Webseiten im Web mit einbezogen werden.

2.3. Wikification via link Co-occurrence

Die Mehrdeutigkeit von Wörtern stellt ein Hauptproblem bei der Identifizierung von Linkkandidaten in Texten dar. Der nachfolgende Ansatz von Cai et al. [CZZW13] beschreibt die Wikifizierung von Texten mit Schwerpunkt auf der Beseitigung von Mehrdeutigkeit von Wörtern.

Der Algorithmus funktioniert in drei Schritten. Zuerst werden die Linkkandidaten des untersuchenden Wikipedia-Artikels anhand des gespeicherten Wikipediadatensatzes untersucht. Dazu wird der zu untersuchende Text in einzelne Teile aufgeteilt. Im zweiten Schritt erfolgt der iterative Prozess nachdem die Mehrdeutigkeit der einzelnen Wörter aufgelöst wird und die eindeutigen Linkkandidaten zu einer Link-Matrix hinzugefügt werden. Es wird solange über den Text iteriert bis keine Mehrdeutigkeit im Text festzustellen ist. Im letzten Schritt wird

[5] http://www.alexa.com/

die gebildete Link-Matrix mit dem geparsten Text verbunden. Die Abbildung 2.2 verbildlicht diesen Prozess.

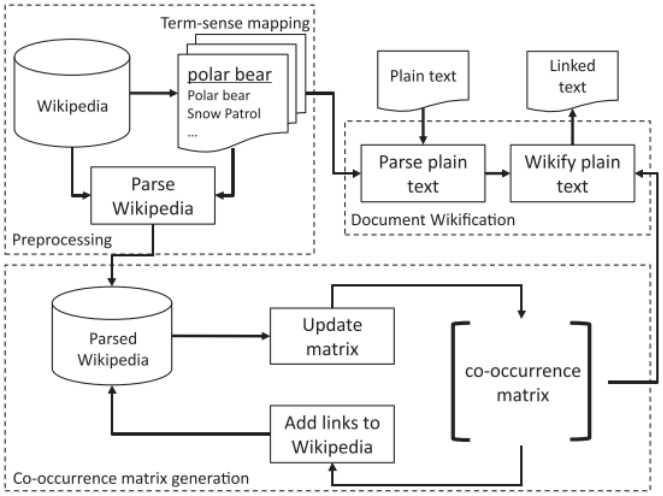

Abbildung 2.2.: Architektur des Wikifizierungs-Frameworks [CZZW13]

Sinnhaftigkeit von Wörtern herausfinden

Um die Bedeutung der Wörter zu finden, werden Bedeutungslisten verwendet. Diese splitten die Sinnhaftigkeit eines Wortes in einzelne Bedeutungen auf. Hierfür werden in der ersten Phase die Links im Wikipedia-Artikel gelöscht. Danach wird der Text in Teile zerteilt und es werden mit Hilfe eines *Natural Language Processing (NLP) Chunkers* die Nominalphrasen extrahiert. In der zweiten Phase werden Linkkandidaten aus den Nominalphasen herausgesucht. Dafür wird geprüft, zu welchen Wörtern ein Wikipedia-Artikel existiert.

Kookkurrenz-Matrix Bildung

Die Aufbereitung der Daten erfolgt in einer *K x K* große Kookkurrenz-Matrix, wo *K* die Anzahl an Konzepten (Sinne) in Wikipedia darstellt. Jedes Element der Matrix beschreibt die totale Anzahl der gemeinsamen Auftritte der zwei verbundenen Konzepte (Sinne) in Wikipedia. Für die Verwendung der Kookkurrenz-Matrix im Algorithmus muss ihre Initialisierung und Anreicherung erfolgen.

Damit die Initialisierung der Matrix erfolgen kann, werden die geparsten Wikipedia Artikel auf die Anzahl der gemeinsamen Auftretenden Konzepte untersucht. Um die Menge an Relationen in der Kookkurrenz-Matrix zu begrenzen und damit Konzepte nicht mit einander verbunden werden, welche semantisch weit auseinander liegen, müssen die zu verbindenden Artikel eine Mindestanzahl an gemeinsamen Verlinkungen W_c aufzuweisen.

Da die so initialisierte Kookkurrenz-Matrix aufgrund der vielen spärlichen Verlinkungen von Wikipedia-Artikeln zu wenige Informationen aufweist, wird sie in dem nächsten Schritt mit weiteren Informationen angereichert. Danach wird die gebildete Bedeutungsliste S_u für einen unverlinkten Kandidaten t_u und die Bedeutungsliste S_i seiner Nachbarn betrachtet. Aufgrund des Satz von Bayes wird die Wahrscheinlichkeit für eine mögliche Verlinkung für einen Linkkandidaten berechnet. Der Wikipedia-Artikel mit der höchsten Warscheinlichkeit wird mit dem Linkkandidaten t_u verbunden. Danach wird die Kookkurrenz-Matrix upgedatet. Das Ganze erfolgt solange bis kein weiterer Linkkandidat mehr gefunden wird.

Untersuchung

Für die Abschlussuntersuchung des entwickelten Verfahrens wurde zuerst der Klassifikator anhand eines Test-Sets von 3.000 der Wikipedia-Artikel aus dem Jahr 2011 mit den größten *PageRank* trainiert. Der Hintergrund davon war, dass Wikipedia-Artikel mit einem großen *PageRank* und damit einer hohen Anzahl an Verlinkungen dazu neigen einen großen Verlinkungsgrad im Text aufweisen.

Um den Vergleich zu anderen entwickelten Wikifizierungs-Verfahren zu ermöglichen, wurden für die Abschlussuntersuchung die Test-Sets aus den Arbeiten von Cucerzan [Cuc07], Kulkarni et al. [KSRC09] und einem von den Autoren erstellten Test-Set bestehend aus 25 Artikeln der *New York Times* und *China Daily* verwendet. Cucerzans Testdaten-Set besteht aus Nachrichtenartikeln und Kulkarnis aus 16.000 extrahierten Wikipedia-Artikeln aus dem Jahr 2008. Der Ansatz von den Autoren konnte mit einem F-Maß von rund 80 % für die untersuchten Testsätze von Cucerzan und Kulkarni et al.,welche unterhalb von einem F-Maß von 65 % geblieben sind, bessere Werte liefern. Die genaue Auflistung der Ergebnisse kann dem Paper entnommen werden.

Fazit

Das vorgestellte Verfahren stellt einen interessanten Ansatz zum Erlernen der Wikifizierung aufgrund eines statistischen Verfahrens dar. Da für die Generierung der Kookkurrenz-Matrix eine große Menge an Datensätzen gehalten werden muss, eignet sich das Verfahren nur bedingt für die Entwicklung der Kontext-Komponente. Dennoch ist die Herangehensweise mit dem Vergleich zu anderen entwickelten Wikifizierungs-Verfahren erstrebenswert. Die Angegebenen Test-Datensätze können so für die Evaluierung des trainierten Klassifikators der Kontext-Komponente verwendet werden.

3. Abschließendes Fazit & Ausblick

Die vorgestellten Ansätze zeigen Verfahren aus der algorithmischen-anwendungsorientierten Perspektive, welche es ermöglichen, Systeme zu konstruieren, die praktisch relevante Lernaufgaben, wie die Generierung von Links als Zusatzinformationen für einen adressierten Aspekt eines Textes, lösen können. Weitere vergleichbare Ansätze finden sich in [MC07, RRDA11, WLWT12]. In all den in der Ausarbeitung vorgestellten Ansätzen wurde überwachtes Lernen verwendet, nachdem der Klassifikator vor dem eigentlichen Einsatz trainiert wurde. Die bearbeiteten Ansätze konzentrieren sich verstärkt auf Wikipedia als Informationsquelle. Es ist denkbar, die gewonnenen Erkenntnisse auf andere Informationsquellen zu projizieren, wenn sie: ein breites Themenfeld abdecken, aus verifizierten Informationen bestehen, semistrukturiert sind und einen einfachen Zugang auf die Daten bieten. Für die Verwirklichung der Kontext-Komponete des IfSS-Systems wurden die erarbeiteten Erkenntnisse aus den vorgestellten Ansätzen im abgeleiteten Projektstrukturplan (siehe A.3) die Arbeitspakete farblich grün hervorgehoben.

Die weiteren Schritte beschäftigen sich mit der Realisierung einer Kontext-Komponente zur Beschaffung von Hintergrundinformationen durch Generierung von Anfragen zu Informationsquellen für Second Screen Anwendungen. Dazu muss ein Klassifikator zum Erlernen der Wikifizierungs-Verfahrens zur Erkennung von Schlagwörtern in Texten trainiert werden. Hierfür soll nach den gewonnenen Erkenntnissen das Klassifikations-Verfahren *Random Forest* verwendet werden. Für das Feature-Set wird eine Kombination der Features aus [MW08] und [OMR13], mit der Erweiterung durch weitere Internet-Metriken wie Aleksa Rank, verwendeten. Für das Training des Klassifikators sollen Wikipedia-Artikel aus Wikipedia gesammelt werden. Das Testen des entwickelten Algorithmus soll anhand von gesammelten Dokumentationsbeschreibungen aus der ZDF-Mediathek erfolgen. Die Evaluierung des erzeugten Klassifikators wird anhand einer quantitativen Analyse durch die Verwendung einer Untersuchungsplattform geschehen. Für den Vergleich der Performance zu anderen entwickelten Wikifikations-Algorithmen soll das Verfahren auf Test-Sets anderer Ansätze angewendet werden.

Aufgrund der Plattformunabhängigkeit wird der IfSS-Client als Web-Applikation konzipiert. Um den Nutzern die Informationen zu liefern die sie interessieren wird eine Auswertungskomponente zur Personalisierung der Suchergebnisse anhand der Nutzerprofile angestrebt. Für eine optimale Präsentation der Inhalte auf einem mobilen Gerät und Ausnutzung der Aufmerksamkeitspanne der Nutzer, sollen die Informationen zusammengefasst werden.

A. Diagramme

A.1. Information for Second Screen - System

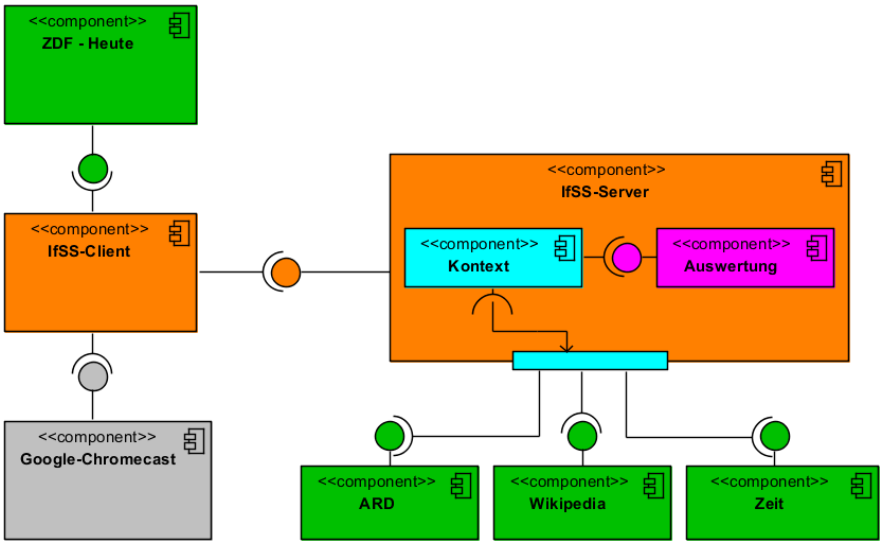

Abbildung A.1.: Die mögliche Realisierung vom IfSS-System als Komponentendiagramm bestehend aus den Hauptkomponenten IfSS-Client und IfSS-Server (Farbe Orange) mit Einbezugnahme von Informationsquellen (Farbe Grün) und Mediastream-Werkzeugen (Farbe Hellgrau).

A.2. Projektstrukturplan - IfSS-Client

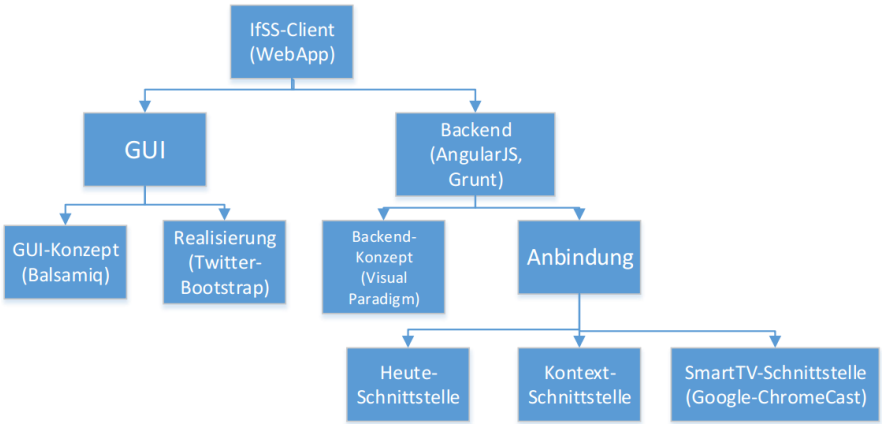

Abbildung A.2.: Projektstrukturplan von den abgeleiteten Arbeitspaketen zur Umsetzung des IfSS-Clients mit Einbezugnahme von Hilfswerkzeugen und Frameworks.

A.3. Projektstrukturplan - Kontext-Komponente

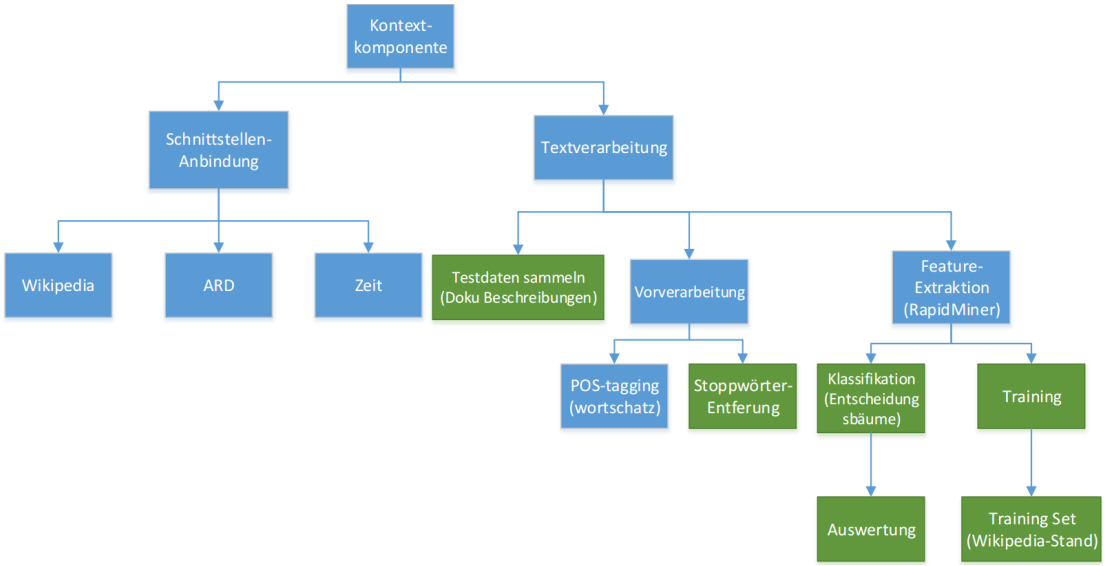

Abbildung A.3.: Projektstrukturplan der abzuarbeitenden Arbeitspakete für die Verwirklichung der Kontext-Komponente, als Unterkomponente des IfSS-Servers, mit der grünen farblichen Hervorhebung von Arbeitspaketen der Verwendung von Erkenntnissen aus den untersuchten Ansätzen.

Literaturverzeichnis

[BRE01] BREIMAN, LEO: Random Forests. In: *Machine Learning* (2001), 5–32. http://download.springer.com/static/pdf/639/art%3A10.1023%2FA%3A1010933404324.pdf?auth66=1407925849_e6ab10aa89873c7f877885aef3949936&ext=.pdf

[CGL+13] Ceccarelli, Diego ; Gordea, Sergiu ; Lucchese, Claudio ; Nardini, Franco M. ; Perego, Raffale: When entities meet query recommender systems. In: *Proceedings of the 28th Annual ACM Symposium on Applied Computing - SAC '13*. New York, New York, USA : ACM Press, März 2013. – ISBN 9781450316569, 933

[Cuc07] Cucerzan, Silviu: Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In: *EMNLP-CoNLL* (2007). http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Large-Scale+Named+Entity+Disambiguation+Based+on+Wikipedia+Data#0

[CZZW13] Cai, Zhiyuan ; Zhao, Kaiqi ; Zhu, Kenny Q. ; Wang, Haixun: Wikification via link co-occurrence. In: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*. New York, New York, USA : ACM Press, Oktober 2013. – ISBN 9781450322638, 1087–1096

[Dem14] Demin, Ivan: Ausarbeitung Master Seminar 1 WiSe 2013 / 2014. (2014)

[DL95] David, C., L. Giroux, S. Bertrand-Gastaldy ; Lanteigne, D.: INDEXING AS PROBLEM SOLVING: A COGNITIVE APPROACH TO CONSISTENCY. In: *In Proceedings of the ASIS Annual Meeting, Medford* (1995), 49–55. http://www.ualberta.ca/dept/slis/cais/david.htm

[EF12] Eimeren, B V. ; Frees, B: Ergebnisse der ARD/ZDF-Onlinestudie 2012. In: *Media Perspektiven* (2012), S. 362–379

[EF13] Eimeren, B V. ; Frees, B: u Multioptionales Fernsehen in digitalen Medienumgebungen. In: *Media Perspektiven* (2013), S. 373–385

[FR09] Ferré, Sébastien (Hrsg.) ; Rudolph, Sebastian (Hrsg.): *Lecture Notes in Computer Science*. Bd. 5548: *Formal Concept Analysis*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2009. http://dx.doi.org/10.1007/978-3-642-01815-2. http://dx.doi.org/10.1007/978-3-642-01815-2. – ISBN 978–3–642–01814–5

[GM04] González, Victor M. ; Mark, Gloria: " Constant , Constant , Multi-tasking Craziness ": Managing Multiple Working Spheres. 6 (2004), Nr. 1, S. 113–120. ISBN 1581137028

[JM09] Jadidinejad, AH ; Mahmoudi, Fariborz: Query wikification: Mining structured queries from unstructured information needs using wikipedia-based semantic analysis. In: *Working Notes for the CLEF 2009 ...* (2009). http://clef.isti.cnr.it/2009/working_notes/Jadidinejad-paperCLEF2009.pdf

[KSRC09] Kulkarni, Sayali ; Singh, Amit ; Ramakrishnan, Ganesh ; Chakrabarti, Soumen: Collective Annotation of Wikipedia Entities in Web Text. (2009), S. 457–465. ISBN 9781605584959

[MC07] Mihalcea, Rada ; Csomai, Andras: Wikify! In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07*. New York, New York, USA : ACM Press, November 2007. – ISBN 9781595938039, 233

[MW08] Milne, David ; Witten, Ian H.: Learning to link with wikipedia. In: *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*. New York, New York, USA : ACM Press, Oktober 2008. – ISBN 9781595939913, 509

[MWR12] Meij, Edgar ; Weerkamp, Wouter ; Rijke, Maarten de: Adding semantics to microblog posts. In: *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*. New York, New York, USA : ACM Press, Februar 2012. – ISBN 9781450307475, 563

[Nie12] Nielsen: *State of the Media - The Cross-Platform Report*. 2012

[Nie14] Nielsen: THE D I G I TA L CONSUMER MEET T O D AY ' S D I G I TA L. (2014), Nr. February, S. 14

[OMR13] Odijk, Daan ; Meij, Edgar ; Rijke, Maarten de: Feeding the second screen: semantic linking based on subtitles. (2013), Mai, 9–16. http://dl.acm.org/citation.cfm?id=2491748.2491751

[PBMW98] Page, Larry ; Brin, Sergey ; Motwani, Rajeev ; Winograd, Terry: 1 Introduction and Motivation 2 A Ranking for Every Page on the Web. (1998), S. 1–17

[RRDA11] Ratinov, Lev ; Roth, Dan ; Downey, Doug ; Anderson, Mike: Local and global algorithms for disambiguation to Wikipedia. (2011), Juni, 1375–1384. http://dl.acm.org/citation.cfm?id=2002472.2002642. ISBN 978–1–932432–87–9

[STB09] Salvucci, Dario D. ; Taatgen, Niels A. ; Borst, Jelmer P.: Toward a unified theory of the multitasking continuum. In: *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*. New York, New York, USA : ACM Press, April 2009. – ISBN 9781605582467, 1819

[TABM03] Trafton, J.Gregory ; Altmann, Erik M. ; Brock, Derek P. ; Mintz, Farilee E.: Preparing to resume an interrupted task: effects of prospective goal encoding and retrospective rehearsal. In: *International Journal of Human-Computer Studies* 58 (2003), Mai, Nr. 5, 583–603. http://dx.doi.org/10.1016/S1071-5819(03)00023-5. – DOI 10.1016/S1071–5819(03)00023–5. – ISSN 10715819

[WLWT12] Wang, Zhichun ; Li, Juanzi ; Wang, Zhigang ; Tang, Jie: Cross-lingual knowledge linking across wiki knowledge bases. In: *Proceedings of the 21st international conference on World Wide Web - WWW '12*. New York, New York, USA : ACM Press, April 2012. – ISBN 9781450312295, 459