

Studienarbeit

Extraktion von räumlichen und zeitlichen Informationen aus Webtexten

Nora Popp

März 2008

Betreuer: Bastian Quilitz, Prof. Ulf Leser, Kathrin Poser

HU Berlin, Institut für Informatik

Inhalt

1	Einleitung.....	3
2	Theoretische Ansätze.....	4
2.1	Methoden zur Extraktion von Geoinformationen.....	4
2.1.1	Web-a-Where: Geotagging Web Content.....	5
2.1.2	Design and Implementation of a Geographic Search Engine	7
2.1.3	STEWARD: Architecture of a Spatio-Textual Search Engine.....	9
2.1.4	A confidence-based framework for disambiguating geographic terms	12
2.1.5	Zusammenfassung	14
2.2	Methoden zur Extraktion von zeitlichen Informationen.....	15
2.2.1	From Temporal Expressions to Temporal Information : Semantic Tagging of News Messages.....	15
2.2.2	Robust Temporal Processing of News.....	17
2.2.3	Zusammenfassung	19
3	Anwendungsbeispiele	19
4	Praktischer Versuch	20
4.1	Extraktion von Ortsangaben	20
4.1.1	Wonach soll gesucht werden?.....	20
4.1.2	Wörterbuch	21
4.1.3	Relative Ortsangaben.....	22
4.1.4	Keine Desambiguierung	23
4.1.5	Methode	23
4.1.6	Korpus.....	24
4.1.7	Ergebnisse.....	25
4.2	Extraktion von Datumsangaben.....	29
4.2.1	Wonach soll gesucht werden?.....	29
4.2.2	Methode	30
4.2.3	Ergebnisse.....	33
5	Abschlussbetrachtung	35
6	Ausblick.....	35
7	Literaturliste.....	37
8	Anhang.....	38

1 Einleitung

Seit der PC in fast alle Firmen und auch viele Privathaushalte eingezogen ist, werden immer mehr Informationen, die früher ausschließlich auf Papier festgehalten wurden, digitalisiert. Dabei handelt es sich sowohl um neue Daten, die fast ausschließlich elektronisch verarbeitet und gespeichert werden, als auch um ältere Daten, die nachträglich digitalisiert werden.

Mit der Verbreitung des Internet ist die Möglichkeit entstanden, viele dieser Daten der Öffentlichkeit oder zumindest bestimmten Benutzergruppen zugänglich zu machen. Im Laufe der Zeit ist das World Wide Web zu einer riesigen Plattform des Datenaustauschs und nicht nur zu einer möglichen, sondern zu einer wesentlichen Quelle der Informationsgewinnung geworden.

Um mit der entstandenen Flut an Daten, die über das Internet erhältlich ist, überhaupt umgehen zu können, und die jeweils relevanten Informationen zu finden, müssen die vorhanden Daten gefiltert werden. Ein besonderes Problem stellt dabei die Vielzahl von unstrukturierten Texten dar. Sie bieten zwar mengenmäßig das größte Informationspotential, sind aber viel schwieriger als strukturierte Daten auf ihren Informationsgehalt hin zu prüfen.

In vielen dieser unstrukturierten Texte finden sich Angaben zu Raum und Zeit, Informationen, die in ganz verschiedenen Problemfeldern von großem Interesse und Nutzen sind. Da die textlichen Informationen jedoch nicht unbedingt offensichtlich relevante Schlagwörter enthalten, ist es oft schwierig bis unmöglich, sie zu finden und zu nutzen. Gerade Ortsangaben können auf sehr unterschiedliche Art und Weise erfolgen, in Form von Koordinaten, Straßennamen oder Städtenamen, von sehr genau beschrieben bis grob umrissen. Und auch Zeitangaben lassen sich nicht unbedingt über Schlagworte in einer Suchmaske finden.

Aus diesem Grund ist es wichtig, Wege zu finden, Informationen über Ort und Zeit aus Texten extrahieren zu können. Sind diese Daten einmal ermittelt, ist es

möglich, sie einem übergeordneten Ort- und Zeitrahmen zuzuordnen und somit nutzbar zu machen.

In dieser Studienarbeit soll zunächst ein Überblick über prinzipielle Möglichkeiten zur Informationsextraktion von Orts- und Zeitangaben aus unstrukturierten Texten gegeben werden. Des Weiteren werden Beispielprojekte vorgestellt, in denen solche Methoden zur Informationsextraktion angewendet werden. Im zweiten Teil der Studienarbeit wird eine Methode zur Extraktion von räumlicher Information und eine Methode zur Extraktion von zeitlicher Information beschrieben und evaluiert.

2 Theoretische Ansätze

2.1 Methoden zur Extraktion von Geoinformationen

Um Fragen wie "Wie werden mögliche Geodaten in einem unstrukturierten Text aus dem Web erkannt und extrahiert?" und "Wie werden mögliche Geodaten desambiguiert?" zu klären, werden im Folgenden vier Projekte vorgestellt, die auf unterschiedliche Weise an diese Themen herangehen. Dabei werden die zugrunde liegenden Daten, die Methoden zur Extraktion von Geodaten und die Methoden zur Desambiguierung der Geodaten dargestellt. Unter Desambiguierung versteht man hier die Auflösung von Mehrdeutigkeiten von geografischen Angaben und die Zuordnung zu einem bestimmten Ort. Zusätzlich werden Begriffe wie Geocoding und der Geofokus einer Website näher erläutert.

Allen Projekten liegt eine ähnliche Idee zu Grunde: die Entwicklung von Suchmaschinen, die die Möglichkeit bieten sollen, Suchanfragen mit Fokus auf bestimmte geografische Regionen stellen zu können.

2.1.1 Web-a-Where: Geotagging Web Content [1]

In [1] wird zwischen zwei verschiedenen geografischen Entitäten unterschieden, die mit einer Internetseite assoziiert werden können, der Quelle und dem Ziel. Die Quelle ist hier der Ursprung der Seite, die physikalische Location des Servers, die Adresse des Autors/Eigentümers, etc.. Das Ziel erschließt sich hingegen aus dem Inhalt der Seite.

Um geografische Entitäten aus dem Inhalt einer Seite zu extrahieren, kann ein einfaches Prinzip angewendet werden: Man vergleicht alle gefundenen Namen auf einer Seite mit einer Liste möglicher Ortsnamen und bestimmt ihre Bedeutung. Das Problem dabei ist, dass ein Großteil der gefundenen Namen mehrdeutig ist. Hierbei wird zwischen geo/non-geo-Ambiguität und geo/geo-Ambiguität unterschieden. Bei der geo/non-geo-Ambiguität hat ein Name sowohl eine geografische, als auch eine nichtgeografische Bedeutung. So kann er z.B. als Nachname oder gewöhnliches Wort gebraucht werden, wie z.B. *London* als Stadtname und als Nachname bei "*Jack London*" oder *Halle* als Stadtname und als "*die Halle*", ein großer Raum. Bei der geo/geo-Ambiguität tragen verschiedene Orte den selben Namen, z.B. *Frankfurt an der Oder* und *Frankfurt am Main*.

Daten

Die Autoren von [1], Amitat et. al., haben mit einem Wörterbuch gearbeitet, das alle Länder und viele ihrer Städte (mit mehr als 5000 Einwohnern), sowie Staaten und Provinzen für die USA, Kanada, Australien, China und UK beinhaltet. Insgesamt sind dort fast 40.000 Orte, mit alternativer Schreibweise und Abkürzungen ca. 75.000 Namen, aufgelistet. Das Wörterbuch bietet einen hierarchischen Blick auf die Welt, der unterteilt ist in Kontinente, Staaten, (Bundes-)Länder und Städte. Dabei wird jede geografische Entität mit einem kanonischen Taxonomieknoten verbunden. Außerdem beinhaltet das Wörterbuch Koordinaten und Bevölkerungszahlen.

Zusätzlich wurde eine spezielle Liste mit Ortsnamen angelegt, die auch als normale Wörter verwendet werden. Dafür wurde gezählt, wie oft ein Wort aus dem Wörterbuch im Korpus vorkommt. Namen, die über 100 mal, aber meistens klein geschrieben vorkommen und Namen, die viel häufiger vorkommen, als es die Bevölkerungszahl erwarten lassen würde, wurden in diese Liste

aufgenommen. Namen aus dieser Liste benötigen eine, nicht näher erläuterte, Bestätigung, um als geografische Referenz gewertet zu werden. Vermutlich geografische Referenzen, die sich in räumlicher Nähe zur eventuellen Referenz befinden.

Geo Extraktion und Desambiguierung

Es wurden alle geografischen Namen von einem Dokument extrahiert, die mittels des Wörterbuches identifiziert werden konnten. Abkürzungen wurden dabei nicht betrachtet.

Der Vorgang der Desambiguierung unterteilt sich in vier Schritte, in denen den gefundenen Namen Vertrauenswerte, Wahrscheinlichkeiten, mit denen ein Name auf einen bestimmten Ort referenziert, zugewiesen werden.

Schritt 1: Der Geotagger weist einem Namen, der eindeutig identifiziert wurde, einen Wert zwischen 0.95 und 1 zu. Dies geschieht, um das hohe Level der Sicherheit anzuzeigen.

Schritt 2: Allen unaufgelösten Namen wird ihr Defaultwert, die Entität mit der größten Bevölkerungszahl, zugeordnet und ein Vertrauenswert von 0.5 zugewiesen.

Schritt 3: Wenn in einem Dokument mehrere Entitäten mit dem selben Namen aufgetreten sind, von denen nur einer qualifiziert wurde, wird die Bedeutung des qualifizierten Namens an die anderen weiterdelegiert. Hierbei wird das Prinzip des "single sense per discourse" angewendet, das davon ausgeht, dass ein Name, der mehrfach erwähnt wird, sich immer auf den selben Ort bezieht, es sei denn, er verweist explizit auf einen anderen Ort. Nach dieser Zuweisung wird ein Vertrauenswert von 0.8 - 0.9 vergeben.

Schritt 4: Alle Namen, deren Ambiguität bisher nicht aufgelöst werden konnte, werden mit Hilfe des Kontexts desambiguiert. Dabei geht man davon aus, das ein Autor immer einen speziellen räumlichen Rahmen intendiert, insoweit er nicht ausdrücklich davon abweicht. Aus diesem Grund versuchen Amitat et. al. hier einen möglichst niedrigen Taxonomieknoten zu finden, der alle unqualifizierten Namen umfasst. Die hierbei zugewiesenen Wahrscheinlichkeitswerte liegen zwischen 0.65 und 0.75.

Dieses Verfahren haben die Autoren auf drei verschiedenen Korpora, aus insgesamt 600 Webseiten mit mehr als 7000 getagten Georeferenzen, getestet. Sie geben dafür eine Precision zwischen 0.63 und 0.82 an. Über die Zahl der gesuchten Namen werden keine Angaben gemacht, somit lässt sich der Recall des Algorithmus nicht bestimmen.

Geofokus

Die Grundidee hierbei ist die, dass, wenn in einem Dokument auf mehrere Orte einer Region referenziert wird, diese Region der Fokus des Dokuments ist. Ein Dokument kann auch mehrere Foki haben.

Um die Wichtigkeit einer Ortsangabe für ein Dokument zu ermitteln, werden den einzelnen Taxonomieknoten des Wörterbuches Werte zugewiesen. Hierbei spielt die Höhe des jeweiligen Knotens eine Rolle. Je genauer eine Ortsangabe in einem Text ist, umso höher ist der zugewiesene Wert. Als Beispiel werden hier die Knoten "Paris/Frankreich/Europa", "Frankreich/Europa" und "Europa" angegeben, wobei der erste Knoten den höchsten Wert erhält. Nach der Aufsummierung aller Knotenwerte für ein Dokument werden die Knoten nach ihrer Wichtigkeit hierarchisch geordnet.

2.1.2 Design and Implementation of a Geographic Search Engine [7]

Geocoding

Unter Geocoding verstehen Markowitz et. al., die Autoren von [7], den Prozess der Zuordnung von geografischen Ortsangaben zu Webpages, die relevante Informationen zu diesen Orten anbieten. Ein Dokument kann mit einer oder mehreren Ortsangaben assoziiert werden und die Sammlung aller Ortsangaben einer Webseite wird der "geografischer Fußabdruck" dieser Webseite genannt. Jeder Ortsangabe in diesem Fußabdruck wird ein Integerwert zugeordnet, der die Gewissheit (Wahrscheinlichkeit) ausdrückt, mit welcher die Webpage relevante Informationen zu diesem Ort anbietet.

Daten

Mit Hilfe des *PolyBot web crawlers* wurden ca. 31 Mio. Webpages von der de-Domain erfasst. Einige Gründe für die Wahl der de-Domain waren das Verständnis der Sprache, der administrativen Geografie und der Konventionen zum Referenzieren geografischer Entitäten. Außerdem sind die *whois*-Einträge für die de-Domain komplett und gut strukturiert. Für die Erkennung geografischer Referenzen wurden zwei Datensätze für Deutschland verwendet. Der erste ordnet jede der 5.000 Telefonvorwahlnummern einer Stadt zu und auch den Koordinaten des Zentrums der Region, die diese Vorwahl umfasst. Der zweite ordnet 82.000 Städten ihre Postleitzahl und ihre Position zu.

Geo Extraktion

Es wurden nur Terme extrahiert, die geografischen Positionen zugeordnet werden konnten, also Städtenamen, Telefonnummern und Postleitzahlen. Zusätzlich wurde die URL der jeweiligen Seite analysiert. Diese Analyse ist jedoch ziemlich schwierig, da einzelne Terme meist nicht separiert sind.

Da viele Terme, die in Städtenamen vorkommen, auch gewöhnliche deutsche oder englische Wörter oder Nachnamen sind, wurde die Menge aller Terme, die in Städtenamen vorkommen, manuell in 3.000 schwache Terme (Terme die auch im normalen Sprachgebrauch vorkommen) und 55.000 starke Terme (solche, die meist nur in Städtenamen gebraucht werden) unterteilt.

Zuerst wurden nur die starken Terme aus den Dokumenten extrahiert, danach wurde nach schwachen Termen gesucht, die in einem vorher festgelegten Abstand von starken Termen auftraten. Schwache Terme, die sich außerhalb dieses Abstandes befanden, wurden ignoriert. Als Beispiel für einen starken Term wurde hier *Frankfurt* genannt. Der schwache Term *Main* wird akzeptiert, wenn er sich irgendwo auf der Seite (Abstand zu *Frankfurt* = ∞) befindet. Der schwache Term *Oder* hingegen, wird nur als zu *Frankfurt* gehörend gewertet, wenn er innerhalb eines Abstandes von 2 vorkommt.

Um die Genauigkeit zu erhöhen, wurden den starken Termen sogenannte *killer terms* und *validator terms* zugeordnet. Wenn ein *killer term* innerhalb eines bestimmten Abstands von einem starken Term vorkommt, wird der starke Term ignoriert. Wenn ein *validator term* nicht innerhalb eines bestimmten Abstands von einem starken Term vorkommt, wird der starke Term auch ignoriert. Zusätzlich

wurde eine Liste mit sogenannten *general killers* angelegt, sie beinhaltet 3.500 allgemeine Vornamen und Titel wie Frau, Herr, Dr. usw..

Desambiguierung

Wie bei Amitay et. al. [1] wird auf das Problem der geo/geo-Ambiguität hingewiesen. Um dieses Problem zu lösen werden zwei Annahmen getroffen. Die erste ist die schon oben genannte "single sense per discourse"-Annahme und die zweite bezieht sich darauf, dass ein Autor zumeist die größte Stadt mit einem bestimmten Namen meint, es sei denn, es gibt starke Indikatoren, die dagegen sprechen.

Um einen Namen einem bestimmten Ort zuzuweisen, werden verschiedene Messwerte angesetzt, z.B. die Anzahl der Worte, die im Dokument gefunden wurden, die in einem Städtenamen vorkommen, das Vorhandensein einer Postleitzahl, die eine Mehrdeutigkeit auflösen kann oder das Auftreten anderer Städtenamen im Dokument, die sich in räumlicher Nähe zum analysierten Namen befinden.

Auf der Grundlage der so erzielten Resultate werden den Ortsnamen Integerwerte zugeordnet, die dann in den geografischen Fußabdruck einer Webseite aufgenommen werden.

In [7] werden keine Angaben über die Güte des Verfahrens angegeben.

2.1.3 STEWARD: Architecture of a Spatio-Textual Search Engine [4]

Daten

Die Autoren von [4], Liebermann et. al., weisen auf das eher kleine Wörterbuch von Web-a-Where [1] hin. Größere Wörterbücher ermöglichen, dass die meisten geografischen Terme aus einem Dokument im Wörterbuch gefunden werden, können jedoch den Tagging-Prozess verlangsamen und reduzieren potenziell die Akkuratheit bei der Desambiguierung.

Sie benutzen für geografische Entitäten der United States das *Geographic Names Information System* (GNIS) mit 2.06 Millionen Einträgen und für nicht-US Entitäten *Geonet Names Server* (GNS) mit 1.08 Millionen Einträgen.

Alle untersuchten Webdokumente wurden nach dem Download standardisiert, durch Nutzung eines ASCII-Formats zur Weiterverarbeitung und einer HTML-Version zu Ansichts- und Annotationszwecken. Wenn möglich, wurden auch Metadaten, wie Titel, Autor, Publikationsdatum und Modifikationsdatum, extrahiert. Die standardisierten Dokumente wurden dann zusammen mit der URL, den möglichen Metadaten, ASCII- und HTML-Version und einer eindeutigen Dokumenten-ID gespeichert.

Geo Extraktion

Zur Extraktion der möglichen geografischen Entitäten wurde hier ein hybrider Ansatz gewählt, der einen *Part-of-Speech* (POS)- und *Named-Entity Recognition* (NER)-basierten Tagger [6] benutzt. Zuerst werden alle Wörter mit ihren korrespondierenden POS-Labels markiert, danach werden nur Eigennamenphrasen, zusammen mit ihrem Kontext, extrahiert. Diese Phrasen werden dann mit dem NER-Tagger bearbeitet. Wenn die Phrase als Ortsangabe markiert wird, wird sie als *Feature* in den sogenannten *Feature-Vektor* des Dokuments aufgenommen, ansonsten wird sie ignoriert. Der *Feature-Vektor* eines Dokuments besteht aus der Menge seiner *Features*.

Für jedes Dokument d erhält jedes Feature f im Feature-Vektor eine eindeutige Feature-ID (FID). Die FID wird zusammen mit dem Start-offset von f in d , der Länge von f , dem Kontext von f und der Dokumenten-ID von d gespeichert.

Nach der Erstellung des Feature-Vektors wird geprüft, ob die enthaltenen Features auf eine geografische Entität referenzieren. Wenn ein Feature im Wörterbuch gefunden wird, werden alle möglichen passenden Einträge (*feature records* von f) extrahiert. Features ohne passenden Eintrag werden aus dem Feature-Vektor gelöscht.

Desambiguierung

Ein Problem, das sich durch die Größe des Wörterbuches ergibt, ist, dass die meisten Features aus d multiple *feature records* haben, auch wenn sie keine geografischen Ortsangaben sind. Deshalb besteht das primäre Ziel darin, jedem Feature f eines seiner *feature records* zuzuweisen.

Eine Beobachtung, die der Algorithmus von Lieberman et. al. ausnutzt ist folgende: Wenn sich ein Autor auf einen relativ unbekanntem geografischen Ort

bezieht, ist es allgemein üblich, Nachbarschaftsreferenzen oder einen hierarchischen Kontext anzugeben. Hier dient die Bevölkerungszahl als Ersatz für die Bekanntheit von Orten.

Der sogenannte "pair strength algorithm" vergleicht Paare von *feature records*, um zu berechnen, ob sie gegenseitig Hinweise aufeinander geben, basierend auf der Bekanntheit jedes Ortes, der Häufigkeit jedes Ortes sowie ihres Dokumentenabstands und des geodätischen Abstands. Als Ergebnis liefert er die "Paarstärke" zweier *feature records*. Der Dokumentenabstand ist hier, gegeben zwei Features f_1 und f_2 , die Differenz vom Offset von f_1 und f_2 vom Beginn des Dokumentes an.

Je höher der Wert der "Paarstärke" eines Paares ist, umso wahrscheinlicher ist es, dass die *feature records* des Paares korrekt sind. Der Algorithmus generiert alle möglichen Paare von *feature records*. Diese werden dann in absteigender Reihenfolge ihrer "Paarstärke" in einer Liste L gespeichert.

Bei jeder Iteration wird das Paar mit höchster "Paarstärke" ausgewählt und aus der Liste entfernt. Damit wird einem oder mehreren Features eines seiner *feature records* zugewiesen, was zu Folge hat, dass einige Paare in L unmöglich werden und deshalb entfernt werden. Wenn L leer ist, wurde jedem Feature eines seiner *feature records* zugewiesen.

Über die Güte des Verfahrens wurden keine Angaben gemacht.

Geofokus

Der geografische Fokus dient als eine Ordnung der geografischen Ortsangaben in einem Dokument. Er wird in absteigender Ordnung nach der Relevanz der Ortsangaben für ein Dokument präsentiert. Als ein einfaches Maß kann die Häufigkeit des Auftretens eines Ortes l in einem Dokument d betrachtet werden. Das Problem dabei ist jedoch, dass jede Ortsangabe isoliert betrachtet wird.

Ein Algorithmus, der containerbasierte oder hierarchische Cluster-Techniken [6] verwendet, hätte den Vorteil, dass die Ortsangaben nach einer natürlichen oder logischen Kategorisierungsmethode sortiert würden, die für Menschen leicht nachvollziehbar wäre. Hierbei besteht das Problem, dass, wenn ein Dokument nur wenige wichtige Ortsangaben enthält, die über eine große Fläche verteilt sind, das Containerobjekt zu groß werden kann, um nützliche Informationen zu liefern. Ein ähnliches Problem sehen Lieberman et. al. darin, dass durch die Benutzung

hierarchischer Knoten, wie in [1], zur Zuordnung keine geografischen Nachbarschaften betrachten werden. Wenn z.B. zwei Städte nahe beieinander liegen, aber durch Ländergrenzen voneinander getrennt sind, werden sie nicht in Zusammenhang gebracht.

STEWARD benutzt deshalb einen Algorithmus, der *Context-Aware Relevancy Determination* (CARD) genannt wird. Zwei Ortsangaben l_1 und l_2 sind kontextuell verwandt in einem Dokument d , wenn sie häufig im Kontext des anderen in d auftreten. Eine Ortsangabe l ist wichtig für d , wenn sie gut über d verteilt ist und kontextuell mit verschiedenen räumlich naheliegenden Orten in d verbunden ist.

2.1.4 A confidence-based framework for disambiguating geographic terms [8]

In [8] beschreiben die Autoren ein System (MetaCarta) zur Desambiguierung geografischer Terme, das nur auf der Zuweisung von Vertrauenswerten, basiert.

Daten

MetaCarta benutzt ein eigenes Wörterbuch, das einige Millionen Namen-Punkte und Namen-Region-Paaren sowie "Umfassungs"-Beziehung zwischen Regionen und Punkten enthält.

Training

Da MetaCarta mit Wahrscheinlichkeiten arbeitet, mit denen ein Name auf einen Ort referenziert, muss die geografische Signifikanz eines Wörterbuchnamens zunächst anhand eines Trainingskorpus berechnet werden. Jedes Namen-Punkt-Paar n,p bekommt eine initiale Wahrscheinlichkeit zugewiesen, die sich aus der durchschnittlichen Wahrscheinlichkeit der Instanz im Trainingskorpus ergibt. Diese initiale Wahrscheinlichkeit wird dann als Ausgangswahrscheinlichkeit für die Namensinstanz bei der Verarbeitung eines Dokuments verwendet.

Geo Extraktion

Für jede potenzielle Georeferenz werden alle bekannten Bedeutungskandidaten identifiziert. Danach wird jeder möglichen Referenz eines Namens n auf einen

bestimmten Punkt p die Wahrscheinlichkeit $c(p,n)$ zugewiesen, mit der n wirklich auf p referenziert.

Desambiguierung

Auch hier wird wieder betont, dass ein Name n auf verschiedene Punkte oder Regionen oder auch auf nicht-geografische Konzepte referenzieren kann. Die Desambiguierung erfolgt mittels lokaler und nicht-lokaler Information innerhalb des Dokuments.

Wortgruppen wie "*city of*" oder "*major of*" vor einem Namen oder Wörter wie "*community college*" nach einem Namen werden als stark positive Indikatoren für die geografische Natur eines Namens betrachtet. Wörter wie *Mr.*, *Dr.* oder normale Vornamen hingegen, werden als stark negative Indikatoren für geografische Bezüge gedeutet.

Mittels Heuristiken wird die Wahrscheinlichkeit $c_{geo}(n)$, dass der Name n auf eine geografische Ortsangabe referenziert, abhängig von der Stärke der Indikatoren modifiziert. Auch andere Hinweise, wie die Abwesenheit von Großbuchstaben im Namen oder die Ähnlichkeit zu einem Akronym, wirken sich auf $c_{geo}(n)$ aus.

Rauch et. al. haben die Beobachtung gemacht, dass es einen hohen Grad von räumlicher Korrelation in geografischen Referenzen gibt, die sich in textlicher Nähe befinden. Dies, und auch die Bevölkerungszahlen von p wirken sich auf die Schätzung der Wahrscheinlichkeit, dass n auf einen Punkt p referenziert, aus.

Zusätzlich, zu den expliziten Referenzen, extrahiert MataCarta auch relative geografische Referenzen, wie "*15 miles northeast of Portland*", und desambiguiert diese in Korrelation zu ihrer Ausgangsreferenz, "*Portland*".

Auch hier wurden keine Angaben zur Güte des Verfahrens angegeben.

Georelevanz

Die Gewichtung eines Ortsnamens n für ein Dokument wird beeinflusst von seiner Position P_n im Dokument und seiner Prominenz B_n . Wobei sich die Prominenz daraus ermitteln lässt, ob n im Titel oder der Überschrift vorkommt, ob n hervorgehoben ist oder in großer Schrift geschrieben wurde, usw.. Die Gewichtungskomponente der Position eines Namens nimmt von einem Maximalwert zu Beginn eines Dokumentes zu einer niedrigen Zahl nahe am Ende

eines langen Dokumentes ab. Am unteren Rand nimmt sie jedoch wieder zu, um der zunehmenden Relevanz von Fußnoten Rechnung zu tragen.

Die Frequenz F_n eines Namens wird auf ähnliche Weise betrachtet, wie bei Standard-Information-Retrieval-Techniken. Außerdem wird die Zahl der anderen geografischen Referenzen S in die Gewichtung einbezogen. Die Gewichtung $E(P_n, B_n, F_n, S)$ multipliziert mit der Geowahrscheinlichkeit $c_{geo}(n)$ ergibt Georelevanz R_g .

2.1.5 Zusammenfassung

Die soeben beschriebenen Methoden zur Extraktion geografischer Information geben einen Überblick über verschiedene Herangehensweisen an dieses Thema.

Allen gemeinsam ist die Verwendung eines Wörterbuchs zur Identifizierung geografischer Entitäten. Während die Autoren von [7] sich auf die Suche nach Information aus einem bestimmten Land (Deutschland) beschränken, werden in [1], [8] und [4] Wörterbücher verwendet, die sehr viel globaler sind. Ein wesentlicher Unterschied liegt in der Größe der Wörterbücher. Das Wörterbuch in [1] ist, mit 40.000 Ortsnamen für Länder der ganzen Welt, im Vergleich zu mehreren Millionen Einträgen in den Wörterbüchern in [4] und [8], sehr klein. Der Nachteil eines kleinen Wörterbuches liegt in der geringeren Anzahl von Georeferenzen, die erkannt werden können. Leider werden in keinem der vier Ansätze Angaben zum Recall der jeweiligen Methode gemacht. Bei der Suche mit einem kleinen Wörterbuch werden wahrscheinlich weniger Namen als mögliche Georeferenzen gewertet, die gar keine sind, als bei großen Wörterbüchern. Je weniger Namen gesucht werden, umso weniger Namen können falsch interpretiert werden. Ob das jedoch als Vorteil angesehen werden kann, ist fraglich. Um die geo-/nongeog-Ambiguität eines Ausdrucks aufzulösen wird in allen vorgestellten Ansätzen der Kontext des Ausdrucks analysiert. Während in [1] bestimmte Terme eine zusätzliche Bestätigung benötigen, um als geografische Referenz betrachtet zu werden, und in [7] einigen Termen *killer terms* und *validator terms* zugeordnet wurden, wird in [4] und [8] überprüft, ob der Kontext eines Terms für oder gegen eine geografische Referenz spricht. Die letzten beiden Ansätze erscheinen flexibler, da sie sich nicht auf bestimmte, vorher festgelegte, geografische Terme

beschränken. Um eventuell auftretende geo-/geo-Ambiguitäten aufzulösen, werden in allen vier Methoden die Nachbarschaftsbeziehungen der gefundenen möglichen Georeferenzen untersucht. Dabei werden Orte, die näher beieinander liegen als wahrscheinlichere Referenzen betrachtet, es sei denn, explizite Angaben sprechen dagegen. Wenn keine Hinweise auf eine bestimmte geografische Referenz gefunden wurden, wird in alle vier Ansätzen davon ausgegangen, dass dann der bekannteste Ort gemeint ist. In [1], [4] und [7] wird die Bekanntheit eines Ortes mit seiner Bevölkerungszahl in Beziehung gesetzt (große Zahl = große Bekanntheit). Da bei MetaCarta [8] mit vorher trainierten Wahrscheinlichkeiten gearbeitet wird, spielen hier Erfahrungswerte eine größere Rolle (bekannter ist, worüber öfter geschrieben wird). Diese Erfahrungswerte hängen jedoch sehr stark vom Trainingskorpus ab.

Nur in [1] wurden Angaben zur Precision der angewendeten Methode gemacht (Werte zwischen 0.63 und 0.82). Da dieser Ansatz, der mit dem kleinsten verwendeten Wörterbuch ist, könnte man davon ausgehen, dass bei den Ansätzen mit größeren Wörterbüchern die Precision, auf Grund der steigenden Anzahl von geo-/nonge- und geo-/geo-Ambiguitäten, sinkt. Da jedoch die Güte der jeweils angewendeten Methoden nicht evaluiert wurde, kann diese Annahme weder bestätigt noch widerlegt werden.

2.2 Methoden zur Extraktion von zeitlichen Informationen

Im Folgenden werden zwei Ansätze vorgestellt, die sich mit der Extraktion von zeitlicher Information aus unstrukturierten Texten befassen.

2.2.1 From Temporal Expressions to Temporal Information : Semantic Tagging of News Messages [9]

In [9] beschreiben F. Schilder und C. Habel ein semantisches Taggingssystem zur Erkennung von zeitlichen Ausdrücken, das für die Bearbeitung deutscher News-Artikel entwickelt wurde. Sie unterscheiden zwischen Zeit-beschreibenden Ausdrücken und Ereignis-beschreibenden Ausdrücken.

Zeit-beschreibende Ausdrücke beinhalten eine zeitliche Information, die mit Hilfe eines Kalender- oder Uhren-Systems angegeben werden kann. Sie werden meistens durch Präpositional-, Adverbial- oder Nominalphrasen ausgedrückt.

Schilder und Habel geben drei verschiedene Arten von zeitlichen Referenzen an: Explizite Referenzen. Das sind Datumsangaben, die sich direkt auf einen Kalendereintrag, z. B. "08.04.01", oder auf eine bestimmte Uhrzeit, als Beispiele werden hier *3 p.m.* und *Midnight* genannt, beziehen.

Indexikalische Referenzen. Hier werden *today*, *by last week* und *next Saturday* als Beispiele angeführt. Es handelt sich hierbei um Ausdrücke, die nur mittels einer Referenzzeit genauer bestimmt werden können, beispielsweise dem Zeitstempel eines Artikels.

Vage Referenzen. Beispiele hierfür sind *in several weeks*, *in the evening*, *by Saturday the latest*. Für solche Ausdrücke ist es schwierig einen genauen Punkt auf einem Zeitstrahl oder einen Zeitintervall anzugeben. Auch in diesen Fällen ist die Extraktion des Zeitstempels eines Artikels wichtig um ihn als Anhaltspunkt auf einem Zeitstrahl zu platzieren und als Bezugspunkt zu nutzen.

Bei Ereignis-beschreibenden Ausdrücken unterscheiden Schilder und Habel zwischen ganzen Sätzen und speziellen Nominalphrasen. In ganzen Sätzen lässt sich die Information über ein Ereignis mit Hilfe des Verbs ermitteln, in Nominalphrasen referenzieren spezielle Nomen auf ein Ereignis, besonders solche, die durch Nominalisierung entstanden sind. Als Beispiel wird die Phrase *after the election* angegeben, wobei *election* hier als zeitlicher Referenzpunkt benutzt wird.

Schilder und Habel betrachten sieben zeitliche Beziehungen ("before", "after", "incl", "at", "starts", "finishes", "excl"), die durch temporale Präpositionen, wie *before*, *on*, *by* explizit gekennzeichnet werden.

Zur Extraktion von zeitlichen Informationen wurde eine Kaskade von finiten Zustandsautomaten (FSTs) benutzt, die mit der Ausgabe eines Part-of-Speech-Taggers gespeist wurden. Der Part-of-Speech-Tagger bestimmt die syntaktische Kategorie und ein Lemma für jedes Eingabewort.

Jeder FST ist auf eine bestimmte Klasse von zeitlichen Ausdrücken spezialisiert. Ein FST, der alle Vorkommen von temporalen Ausdrücken, die eine Zeitangabe kennzeichnen, markiert, benutzt die Informationen des Part-of-Speech-Taggers und eine vordefinierte Klasse von temporalen Lemmas (Wochentage,

Monatsnamen, generelle Beschreibungen wie *Mittag, Woche, Jahr* etc.). Mit diesem Automaten können auch zusammengesetzte Wörter wie *Rekordjahr* und *Osterferien* erkannt werden. Ein anderer Automat extrahiert den Zeitstempel eines Dokuments, der für die Auflösung indexikalischer und vager Referenzen nötig ist. Ein weiterer Automat markiert alle Verben, da diese temporalen Ausdrücke ein Ereignis beschreiben.

Der Tagger wurde auf einem Korpus von 10 News-Artikeln der Financial Times Deutschland getestet. Schilder und Habel geben für einfache temporale Ausdrücke eine Precision von 92.11 und einen Recall von 94.09 an. Für komplexe temporale Ausdrücke, Präpositionalphrasen, wird eine Precision von 87.30 und ein Recall von 90.66 angegeben.

2.2.2 Robust Temporal Processing of News [5]

I. Mani und G. Wilson beschreiben in [5] eine weitere Methode zur Auflösung temporaler Ausdrücke. Besondere Aufmerksamkeit erhalten dabei indexikalische Ausdrücke, wie *now, today, tomorrow, two weeks ago* etc..

Zeitpunkte werden hier als Einträge eines kalendarischen Koordinatensystems, unter Benutzung des ISO-Standard CC:YY:MM:DD:HH:XX:SS, betrachtet. Ein zeitlicher Ausdruck bestimmt also einen genauen Punkt auf dem Zeitstrahl. Zeitausdrücke, die ein Intervall beschreiben, als Beispiel wird hier unter anderem "*from 3 pm to 6 pm*" genannt, werden als zwei einzelne Ausdrücke repräsentiert. Intervallausdrücke, die keinem bestimmten Zeitpunkt zugeordnet werden können, wie "*half an hour (long)*" oder "*(for) one month*" werden nicht ausgewertet. Ausdrücke, die sehr allgemein und unspezifisch sind, wie "*April*" in "*April is usually wet*" oder Indefinite wie "*a Tuesday*", werden ohne Zuweisung eines Zeitpunktes getagt.

Bevor ein Dokument mit dem Temporal-Tagger bearbeitet werden kann, wird es in Sätze und Worte zerlegt und mit einem Part-of-Speech-Tagger analysiert. Der Temporal-Tagger ist in drei Module unterteilt. Das erste identifiziert Zeitausdrücke als solche. Ein zweites Modul löst unabhängige Zeitausdrücke auf und weist ihnen ihre Werte zu. So erhält beispielsweise der Ausdruck "*June 1999*" den Wert 19:99.06. Anschließend löst das dritte Modul kontextabhängige

Zeitangaben, wie indexikalische, auf. Dafür wird eine Referenzzeit benötigt, auf die der Ausdruck bezogen wird. Diese Referenzzeit ist entweder das Datum der Entstehung eines Dokuments oder der *temporale Fokus* (die Zeit, über die im Dokument gesprochen wird). Um zeitliche Ausdrücke, die nur mit Hilfe des Kontexts einem Zeitpunkt zugewiesen werden können, auszuwerten und ihnen Werte zuzuweisen, wird eine geordnete Liste von Regeln verwendet. Dabei werden folgende Fälle betrachtet:

Ausdrücken wie "yesterday", "today" oder "tomorrow" wird ein Wert zugewiesen, der einen bestimmten Abstand von der Referenzzeit hat.

Bei Ausdrücken wie "next month" oder "last year" werden die lexikalischen Markierungen (*next*, *last*) verwendet, um die Richtung und die Größe des Abstandes zur Referenzzeit zu bestimmen.

Um Ausdrücke wie *Thursday* in "the action taken Thursday" oder einfach nur "February" aufzulösen, wird nach Verben im selben Satz gesucht. Je nach der Zeitform des Verbs (Vergangenheit, Präsens oder Zukunft) wird der Abstand und dessen Richtung zur Referenzzeit ermittelt.

Wenn einem Ausdruck bis dahin kein Wert zugewiesen konnte, wird nach zusätzlichen lexikalischen Markierungen wie "since" oder "until" gesucht, da diese eine Richtung für den zeitlichen Abstand angeben.

Wenn für eine Datumsangabe keine Richtung von der Referenzzeit ermittelt werden kann, weil zum Beispiel die Jahreszahl fehlt, wird dieser Ausdruck nicht als Zeitangabe getagt, es sei denn, das Datum hat einen geringeren Abstand als einen Monat zur Referenzzeit, dann wird dem Ausdruck diese Jahreszahl zugeordnet.

Da bei einem Ausdruck wie zum Beispiel "today" zwei verschiedene Lesarten möglich sind (der Bezug auf den heutigen Tag oder das allgemeinere *heutzutage*) wurden zu den per Hand erstellten Regeln zusätzlich Regeln mit Hilfe von maschinellem Lernen erstellt.

Mani und Wilson haben ihre Methode auf einem Korpus von 22 Print-News-Artikeln der New York Times und 199 Transkriptionen von Rundfunknews getestet. Sie erzielten dabei eine Akkuratheit von 83.2% (F-measure).

2.2.3 Zusammenfassung

In den beschriebenen Methoden zur Extraktion von zeitlicher Information wird zwischen verschiedenen Ausprägungen von Repräsentation von Zeitangaben unterschieden. Die Angaben zur Extraktion von expliziten Zeitangaben sind in beiden Ansätzen nur grob umrissen. In [5] wurden jedoch Möglichkeiten erläutert, wie indexikalische Zeitausdrücke, mit Hilfe von Referenzzeiten, einem genauen Punkt auf dem Zeitstrahl zugeordnet werden können.

Die Ergebnisse der beiden Methoden lassen sich nur schwer miteinander vergleichen. Die hohen Werte für Precision und Recall in [9] wurden auf einem Testkorpus aus nur 10 News-Artikeln der Financial Times Deutschland erreicht. Auch wenn in diesen Artikeln sehr viele Zeitangaben vorkommen, kann man annehmen, dass die Ausprägungen der Zeitangaben innerhalb einer Zeitung, in nur 10 Artikeln, nicht sehr variieren. Das Ergebnis in [5] wurde dagegen auf einem Korpus aus 221 Texten aus zwei verschiedenen Quellen erzielt.

3 Anwendungsbeispiele

Die in Abschnitt 2 vorgestellten Methoden ermöglichen die Extraktion von Angaben zu Raum und Zeit aus unstrukturierten Texten. Diese Angaben sind in ganz verschiedenen Problemfeldern von großem Interesse und Nutzen. So ist es zum Beispiel beim Katastrophenmanagement von großer Wichtigkeit, so genaue Informationen wie möglich über den Ort und die Zeit eines Unglücks oder einer Maßnahme zu erhalten, um dann schnell und effektiv reagieren zu können.

Fahland et. al. [3] haben mit HUODINI einen Prototyp eines System zur flexiblen Integration und Visualisierung heterogener Datenquellen für das Katastrophenmanagement entwickelt. Im Laufe der vergangenen Jahre hat sich gezeigt, dass, im Falle einer Katastrophe, private Internetnutzer oft sehr schnell Informationen dazu in Form von Texten und/oder Bildern veröffentlichen. Diese Informationen sind meist sehr aktuell, da sie nicht erst über den Umweg einer Nachrichtenzentrale oder andere öffentliche Anlaufstellen verfügbar gemacht werden müssen. Um die geografischen Information aus solchen Texten nutzbar

machen zu können, müssen sie zunächst als solche erkannt und extrahiert werden. HUODINI verfolgt hierbei einen Ontologie-basierten Ansatz und benutzt ein Wörterbuch mit relevanten Termen, um diese in den Texten zu identifizieren.

Eine anderes Beispiel für die Verarbeitung geografischer Information beschreiben Chen et. al. in [9]. Mit GeoTracker haben sie, unter anderem, ein System für geospatiales Browsen von News entworfen. Hierbei werden aus den Newsartikeln explizite Ortsangaben mit Hilfe eines regelbasierten Taggers extrahiert. Mittels dieser Ortsangaben werden die Artikel bestimmten Orten auf einer Landkarte zugeordnet. Somit kann nach regionsbezogenen Artikel gesucht werden. Außerdem kann der Benutzer News von einem beliebigen (vergangenen) Zeitpunkt auswählen.

4 Praktischer Versuch

4.1 Extraktion von Ortsangaben

Da der Begriff "Geoinformation" ein sehr weit gefasster ist, muss zunächst festgelegt werden, nach was für Informationen genau gesucht werden soll. Geografische Information, also Information, die sich auf eine geografische Entität bezieht, kann zum Beispiel durch den Namen der geografischen Entität (z.B. "*Berlin*") oder deren Koordinaten (Länge:13.4/13°24'0",Breite:52.5167/52°31'0") oder eine Beschreibung ("*Hauptstadt von Deutschland*") angegeben werden.

4.1.1 Wonach soll gesucht werden?

Im Rahmen dieser Studienarbeit soll nach allen Entitäten gesucht werden, die sich namentlich auf einen Ort innerhalb Deutschlands beziehen. Zu den gesuchten Orten gehören Bundesländer, Regionen, Städte, Dörfer, Flüsse, Seen, Berge und Deutschland als Land.

Wenn in einem Text der Ausdruck "in *Hamburg*" vorkommt, ist das eine explizite Referenz auf einen Ort namens *Hamburg*. Wenn in einem Text der "Bürgermeister von *Hamburg*" erwähnt wird, ist auch das eine explizite Nennung des Namens *Hamburg*, der den Wirkungsbereich des Bürgermeisters beschreibt. Dasselbe wird mit "*Hamburgs* Bürgermeister" ausgedrückt und beinhaltet somit auch die selbe geografische Information. Der Ausdruck "*Hamburger* Bahnhof" bezieht sich ebenfalls namentlich auf die geografische Entität *Hamburg*, auch wenn damit noch nicht gesagt ist, ob es sich dabei um den Bahnhof von *Hamburg* handelt, oder um eine Einrichtung deren Namen sich auf *Hamburg* bezieht. Diese Einrichtung kann sich durchaus auch an einem anderen Ort als Hamburg befinden. Bei der Suche nach Information bezüglich einer geografischen Entität wird in dieser Arbeit nicht unterschieden, ob der Ortsname als solcher explizit erwähnt wird, oder ob er zur Beschreibung anderer Dinge dient.

4.1.2 Wörterbuch

Um Ortsangaben auch als solche identifizieren zu können, bedarf es einer Referenz, auf die sich bezogen werden kann. Dafür wurde hier die DE-Datei von *GeoNames* [10] als Wörterbuch herangezogen (Stand Januar 2008). Diese Datei beinhaltet 174.976 Einträge von deutschen geografischen Entitäten, davon haben 9.364 Einträge zusätzlich mindestens einen Alternativnamen (Gesamttokenzahl: 232.708 ; durchschnittliche Tokenlänge: 8,4 Zeichen; Standardabweichung: 3,16 Zeichen). Für jeden Eintrag werden Längen- und Breitengrade, sowie, wenn möglich, Bevölkerungszahlen angegeben. Die Zahl der Einträge lässt nicht unbedingt auf 174.976 verschiedene geografische Referenzen schließen. Für viele Orte gibt es mehrere Einträge mit unterschiedlichen Längen- und Breitengraden, unterschiedlichen Bevölkerungszahlen oder auch mit zusätzlichen Angaben im Namen, wie z.B. "*Frankfurt*", "*Frankfurt am Main*" und "*Frankfurt am Main, Stadt*".

Diese Daten werden als SQL-Tabelle gespeichert.

4.1.3 Relative Ortsangaben

Oft werden Ortsangaben nicht nur durch die bloße Nennung eines Namens gemacht. Häufig wird die geografische Lage eines Ortes auch relativ zu anderen Orten beschrieben, zum Beispiel durch Ausdrücke wie "im Norden von *Berlin*", "südöstlich von *Berlin*", "in der Nähe von *Berlin*", "bei *Berlin*" u.s.w. Um solche relativen Ortsangaben zu identifizieren, wurde eine Sonderliste mit entsprechenden Ausdrücken angelegt. Solch eine Liste kann allerdings nur einen Ausschnitt aller möglichen relativen Beschreibungen wiedergeben. Da sowohl deutsche als auch englische Texte untersucht werden sollen, wurden auch entsprechende englischsprachige Ausdrücke in die Liste aufgenommen (Abb.1)

Abb. 1: Sonderliste für relative Ortsbeschreibungen

in, im, auf, bei, am, von,
at, on, of, by,
der, die, das, the,
Altstadt, city, river, River,
Kreis, area, Nähe, nahe, near,
Norden, Osten, Süden, Westen,
north, east, south, west,
nördlich, östlich, südlich, westlich,
northern, eastern, southern, western,
Südosten, Südwesten, Nordwesten, Nordosten,
southwest, southeast, northwest, northeast,
südöstlich, südwestlich, nordwestlich, nordöstlich,
southwestern, southeastern, northwestern, northeastern

4.1.4 Keine Desambiguierung

Trotz der Längen- und Breitenangaben im Wörterbuch wird in dieser Arbeit auf den Versuch, Desambiguitäten aufzulösen, verzichtet. Wenn also ein Ortsname mehrfach in Deutschland vergeben ist, dann geht es hier nur darum, eine Referenz auf den Namen in einem Text zu identifizieren, und nicht, ihn bestimmten Koordinaten innerhalb des Landes zuzuordnen. Ähnlich verhält es sich bei beschreibenden Ausdrücken wie "*Hamburger* Bahnhof". Hier sollte die Entität *Hamburg* entdeckt werden, ohne Rücksicht auf die wirkliche Lokalisierung des Bahnhofs.

4.1.5 Methode

Zur Extraktion von geografischen Informationen wurde eine einfache Methode in Java implementiert. Die Grundidee dabei ist, zu testen, wie gut oder schlecht die Ergebnisse durch einen einfachen Stringvergleich sind. Als Vergleichsstrings dienen hier die Einträge des Wörterbuches.

Um die Dokumente zu bearbeiten, werden sie in Token zerlegt. Dabei wird hier als Tokentrenner einfach das Leerzeichen verwendet. Da lediglich ein Stringvergleich vorgenommen werden soll und Satzstrukturinformationen hier nicht beachtet werden, ist eine Analyse mittels eines Part-of-Speech-Taggers nicht nötig.

Damit ein Token mit einem Eintrag im Wörterbuch oder einem Eintrag der Sonderliste identifiziert werden kann, werden zunächst alle umgebenden Zeichen, die keine Buchstaben sind entfernt. Zum Beispiel wird das Token "*(Berlin)*," zu "*Berlin*".

Danach wird getestet, ob das "saubere" Token in der Sonderliste vorkommt. Ist das der Fall, wird es zwischengespeichert und das nächste Token wird für diesen Test angefordert. Ist das nicht der Fall, wird das Token zum nächsten Test weitergereicht.

Hier wird nun geprüft, ob das Token, als ganzer Eintrag oder als Teilbegriff am Anfang eines zusammengesetzten Ausdrucks, im Wörterbuch, der Liste der Namen oder der Alternativnamen, vorkommt. Als Teilbegriff wird hier kein

Substring, sondern nur ein ganzes Wort, akzeptiert, z.B. 'Frankfurt' in 'Frankfurt Oder', nicht aber "Frank". Groß- und Kleinschreibung wird unterschieden. Wenn das Token nicht im Wörterbuch gefunden wird, werden alle zwischengespeicherten Token, die in der Sonderliste gefunden wurden, verworfen und die Suche beginnt mit dem nächsten Token von vorn. Wird das Token gefunden, dann wird das nächste Token angefordert und mit dem vorherigen Token als zusammengesetzter Ausdruck im Wörterbuch gesucht. Die Sonderliste wird in diesem Fall nicht mehr durchsucht, da davon ausgegangen wird, dass sich Ausdrücke wie "nördlich von" oder "near by" vor einer Ortsangabe befinden.

Solange die Suche im Wörterbuch ein positives Ergebnis liefert, wird immer das folgende Token an den Suchbegriff angehängt. Sobald kein passender Eintrag im Wörterbuch gefunden wird, wird der, eventuell zusammengesetzte, Ausdruck ohne das zuletzt angefügte Token, aber mit den eventuell gespeicherten Token aus der Sonderliste, als Ergebnis der Suche geliefert. Das gilt jedoch nur, wenn der Ausdruck als ganzer Eintrag im Wörterbuch gefunden wurde. Wenn der gefundene Ausdruck nur ein Teilbegriff eines Wörterbucheintrages ist, werden die Token verworfen.

Das zuletzt betrachtete Token wird dann als einzelnes dem Algorithmus wieder zugeführt. (Beispiel in Abb. 2)

4.1.6 Korpus

Um die entwickelte Methoden auf Texten aus dem Web zu testen, wurde ein Testkorpus erstellt. Es besteht aus 64 deutsch- und englischsprachigen Dokumenten, die von flickr.com bezogen wurden. flickr.com ist hauptsächlich eine Plattform für Fotos, aber viele dieser Fotos sind mit Beschreibungen versehen.

Da sich die Untersuchung auf geografische Entitäten innerhalb Deutschlands bezieht, wurde eine flickr-Suche mit entsprechenden Schlagworten gestartet. Die Schlagworte waren alle Einträge des Wörterbuches, die eine Bevölkerungszahl von mehr als 50.000 Einwohnern aufweisen, insgesamt 761 Einträge. Die Texte der Resultate beinhalten nicht zwangsläufig den jeweiligen Suchbegriff, da dieser auch in den Metainformationen eines Dokuments enthalten sein kann. Da bei der

Abb. 2: Beispiel für Extraktion von Ortsangaben

Beispiel: "... north of the Thüringer Wald, east of ..."		
		Speicher: ""
north	→ Test Liste : ok	Speicher: "north"
of	→ Test Liste : ok	Speicher: "north of"
the	→ Test Liste : ok	Speicher: "north of the"
Thüringer	→ Test Liste : negativ	
Thüringer	→ Test Wörterbuch : ok	
Thüringer Wald	→ Test Wörterbuch : ok	
Thüringer Wald east	→ Test Wörterbuch : negativ	
<u>Ergebnis:</u> "north of the Thüringer Wald"		Speicher: ""
east	→ Test Liste ...	
...		

automatisierten flickr-Suche pro Anfrage nur eine begrenzte Anzahl von Ergebnissen geliefert werden, entstand zunächst ein Korpus mit 7313 verschiedenen Dokumenten. Viele der Texte waren jedoch extrem kurz, sie bestanden oft nur aus wenigen Worten. Etwa 300 Dokumente hatten eine Wortanzahl von über 100 Worten. Nachdem alle nichtdeutschen und nichtenglischen Texte und Texte mit zu vielen Sonderzeichen aussortiert wurden, blieben 64 Texte mit insgesamt 12710 Token übrig. Von den 64 Texten sind 36 Texte deutschsprachig, 14 Texte englischsprachig und 14 Texte enthalten sowohl deutschsprachige als auch englischsprachige Abschnitte.

4.1.7 Ergebnisse

Im Testkorpus wurden durch manuelle Suche insgesamt 669 Ausdrücke mit geografischer Information gefunden, die den oben genannten Anforderungen entsprechen. Die Evaluierung der Ergebnisse wurde ebenfalls manuell durchgeführt.

Von den 669 erwarteten Ausdrücken wurden durch die Anwendung des Algorithmus 510 Ausdrücke gefunden. Relative Angaben wurden, solange sie in der Sonderliste aufgeführt waren, dabei korrekt erkannt. Bei einem Ausdruck wie "*East Berlin*" wurde zwar *Berlin* erkannt, nicht jedoch der Zusatz *East*, da dieser nicht in der Sonderliste steht. Solch ein Ergebnis wurde jedoch auch als richtig-erkannt gewertet, da sich die hauptsächliche Suche auf den konkreten Ort bezieht. In Tabelle 1 sind die Ergebnisse für die deutschen, englischen und zweisprachigen Texte aufgeführt. Die Einzelergebnisse für jedes Dokument sind im Anhang angegeben.

Tab. 1: Ergebnisse der Extraktion von geografischen Referenzen

	alle Texte (64)	dt. Texte (36)	engl. Texte (14)	dt./engl. Texte (14)
gesucht	669	367	91	211
true pos.	510	256	82	172
false neg.	159	111	9	39
false pos.	489	347	30	112

Bei den 159 nicht gefundenen Ausdrücken (false negatives) handelt es sich in den meisten Fällen um Ausdrücke wie "*Berliner Dom*" oder "*eine Stadt Deutschlands*". Solche Wendungen sind im Deutschen sehr gebräuchlich, können aber nicht mit Hilfe des Wörterbuches erkannt werden, da dort nur die Normalformen der Ortsnamen aufgelistet sind. Der Anteil nichtgefundener Ausdrücke ist in den rein englischsprachigen Texten sehr viel kleiner. Ausdrücke wie "*Germany's*" kamen in den rein englischsprachigen Texten, bei 91 gesuchten Ausdrücken, nur viermal vor.

Ein kleinerer Teil der nicht gefundenen Ausdrücke erklärt sich aus der Tatsache, dass es sich dabei um historische Ortsbezeichnungen handelt, die nicht im Wörterbuch aufgeführt sind.

Da bei der Suche nach geografischen Referenzen zwischen Groß- und Kleinschreibung unterschieden wurde, wurden kleingeschriebene Ausdrücke nicht gefundenen. In Tabelle 2 wird der Recall der Methode angegeben.

Tab. 2: Recall bei Extraktion geografischer Referenzen

	alle Texte (64)	dt. Texte (36)	engl. Texte (14)	dt./engl. Texte (14)
Recall	0.76	0.70	0.90	0.82

Zusätzlich zu den richtigen Ergebnissen hat der Algorithmus jedoch auch 489 Ausdrücke als geografische Referenzen gewertet, die keine waren (false positives). Einen kleinen Auszug der falsch erkannten Referenzen zeigt Abb. 3.

Abb. 3: Beispiele falsch erkannter Referenzen

<p>Eine, Wetter, Mark, Weg, Sie, Klaus, Stadt, Schule, Schülern, Burg, Höhe, Mitte, Hase, Igel, Kloster, Städten, Apostel, Buch, Alten, Bad, Freiheit, Welt, Kern, Kahn, Sepp, Maier, Japan, USA, Helden, Thomas, Bahnhof, Wall, Fall, Haus, Wilhelm, Dom, Forster, Ort, Tal, Ausbau, Marx, Müller, Vogel, Bau, Kaiser, Quer, Langhaus, Neubau, Kirche, Häuschen, Winter, Schnee, Mühe, Wolfgang, Freundschaft, Quellen, Siedlung, Wasser, Wiesen, Dies, Grund, König, Truppen, Ludwig, Wolf, Schinkel, Fuß, Länge, Breite, Ecke, Hauptquartier, Baumeister, Friedrich, Maximilian, Kalle, Albert, Franz, Herder, Bauhaus, Boden, Markt, Besitz, Bach, Abel, Abtei, Königshof, Alte, Berge, Karl, Mayer, May, Grimm, Hospiz, Schlacht, Schloß, See, Adler, Höh, Geist, Ob, Grauen, Rampe, Brunnen, Düne, Strände, Garten, Herzogs, Dauer, Jägers, Stein, Linde, Hügel, Hof, Blick, Weinbergen, Höfe, Häuser, Stelle, Hammer, Hütte, Mühle, Wind, Mühlen, Hunger, Max, Kaspar, Schmid, Kurt, Lippe, Dorf, Metternich, Heinz, Ende, Straße, Most, Hall, Peter, Reisen, Wegen, Maler, Dreieck, Grenze, Anbau, Mauer, Lage, Sonne, Lehr, Koch, Feld, Rathaus, Zuschlag, Strassen, Gassen, Wohnen, Schau, Wort, Zweck, Grube, Frieden, Ruhe, Kasernen, Scheunen, Handwerk, Strand, Weide, Wache, Wolken, Grün, Welle, Lauf, Lust, Grab, Blume, Hafen, Häusern, Universität, Aula, Kirchen, Macken, Straßen, Strenge, Trost, Meine, Bäcker, Bank</p>
--

Angesichts dieser Ergebnisse kann man Zweifel an der Qualität des Wörterbuches bekommen, aber ein Blick in den Autoatlas verrät, dass viele dieser Ausdrücke durchaus Ortsbezeichnungen sein können.

Aus den 36 rein deutschsprachigen Texten wurden insgesamt 347 Ausdrücke fälschlicherweise extrahiert, aus den 14 rein englischsprachigen Texten dagegen nur 30. Trotz der unterschiedlichen Textanzahl wird deutlich, dass das Problem hier hauptsächlich bei den deutschsprachigen Texten liegt. Das hat verschiedene Ursachen. Bei den meisten falsch erkannten Ausdrücken aus den englischen Texten handelt es sich um Personennamen. Es wurden aber auch englischsprachige, nichtgeografische, Ausdrücke extrahiert, die, zum Beispiel am Satzanfang, großgeschrieben wurden. Das deutet darauf hin, dass bei einer Suche, bei der nicht zwischen Groß- und Kleinschreibung unterschieden wird, auch bei englischsprachigen Texten mehr Ausdrücke fälschlich als geografische Referenzen gewertet werden würden. Bei den meisten falsch erkannten Ausdrücken aus den deutschsprachigen Texten handelt es sich um Wörter des allgemeinen Sprachgebrauchs. Das liegt unter anderem daran, dass Ortsnamen aus Namen für Dinge des alltäglichen Lebens abgeleitet wurden. Das legt die Annahme nahe, dass bei einer Suche in englischen Texten nach geografischen Referenzen in englischsprachigen Gebieten, ähnliche Probleme auftreten, wenn nicht zwischen Groß- und Kleinschreibung unterschieden wird.

An dieser Stelle wird sehr deutlich, wie wichtig eine Auswertung des Kontexts von Ausdrücken ist.

In Tabelle 3 wird die Precision der implementierten Methode angegeben.

Tab. 3: Precision bei Extraktion geografischer Referenzen

	alle Texte (64)	dt. Texte (36)	engl. Texte (14)	dt./engl. Texte (14)
Precision	0.51	0.42	0.73	0.60

Diese Resultate lassen nicht wirklich mit den zu Beginn vorgestellten Methoden vergleichen. Nur Amitat et al. [1] haben die Precision (0.63–0.82) ihres Algorithmus angegeben. Allerdings wurden diese Ergebnisse auf insgesamt 600,

meist englischsprachigen, Webseiten mit mehr als 7000 getagten Georeferenzen, erzielt. Aufgrund der festgestellten Unterschiede in den Ergebnissen, bei deutschen und englischen Texten, könnten nur die Ergebnisse der englischen Texte zum Vergleich herangezogen werden. Die hier untersuchten 14 englischsprachigen Texte bilden jedoch kein repräsentatives Vergleichskorpus.

4.2 Extraktion von Datumsangaben

Wie bereits bei Schilder & Habel [9] und bei Mani & Wilson [5] erwähnt, kann zwischen expliziten und relativen Zeitangaben unterschieden werden. Explizite Zeitangaben sind solche, die sich auf einen genauen Punkt in einem kalendarischen System beziehen. Relative Zeitangaben hingegen benötigen einen Bezugspunkt, von dem aus ihre Position im kalendarischen System berechnet werden kann.

4.2.1 Wonach soll gesucht werden?

Im Rahmen dieser Arbeit werden nur explizite Datumsangaben betrachtet. Da es sowohl im Deutschen als auch im Englischen eine Vielzahl verschiedener Formate für die Angabe eines Datums gibt, müssen zunächst Muster festgelegt werden, nach denen gesucht werden soll. In diesem Fall wird nach allen Datumsangaben gesucht, die durch Zahlen und ".", z.B. *01.01.2008*, sowie durch Monatsnamen, z.B. *1. Januar 2008*, gekennzeichnet sind. Zusätzlich soll nach Wochentagen und "Jahrhundert"-Angaben gesucht werden. Um Monatsnamen, Wochentage und "Jahrhundert"-Angaben zu identifizieren wurde eine Wortliste (Abb. 4) angelegt. Außerdem wurden dort Ausdrücke wie "1st", "2nd" usw. aufgenommen, um Ausdrücke wie "1st century" zu identifizieren.

Abb. 4: Wortliste für Extraktion von Datumsangaben

```
Montag, Dienstag, Mittwoch,  
Donnerstag, Freitag, Sonnabend, Sonntag,  
Monday, Tuesday, Wednesday,  
Thursday, Friday, Saturday, Sunday,  
Januar, January, Februar, February,  
März, March, April, Mai, May, Juni, June,  
Juli, July, August, September, Oktober, October,  
November, Dezember, December,  
Jahrhundert, Jahrhunderts, Jh, century,  
1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th, 9th, 10th,  
11th, 12th, 13th, 14th, 15th, 16th, 17th, 18th, 19th, 20th,  
21st, 22nd, 23rd, 24th, 25th, 26th, 27th, 28th, 29th, 30th,  
31st
```

4.2.2 Methode

Zur Extraktion von Datumsangaben wurde eine Methode in Java implementiert, die oben genannte Formate, durch Erkennung bestimmter Muster, identifizieren soll. Um die Dokumente zu bearbeiten, werden sie, wie bei der Extraktion von Geodaten, in Token zerlegt. Dabei wird auch hier als Tokentrenner das Leerzeichen verwendet.

Damit ein Token mit einem Eintrag in der Wortliste oder einem der gesuchten Muster identifiziert werden kann, werden zunächst alle umgebenden Klammern und Kommata entfernt. Zum Beispiel wird das Token "(März)," zu "März". Danach wird getestet, ob das "reine" Token mit einem Buchstaben beginnt. Ist das der Fall, werden eventuell vorkommende Punkte am Ende des Tokens entfernt, zum Beispiel, wenn das Token an einem Satzende steht. Anschließend wird geprüft, ob das Token in der Wortliste enthalten ist. Ist das Token in der Liste, wird es zwischengespeichert, wenn nicht, wird es verworfen. Danach wird das nächste Token angefordert.

Wenn das betrachtete Token mit einer Zahl beginnt, werden zunächst folgende Fälle geprüft:

Tokenlänge 6 : Das Token wird akzeptiert und zwischengespeichert, wenn es die Form 'xx.xx.' hat, wobei x hier für eine Zahl steht, sonst wird es verworfen.

Hier werden eventuelle Angaben von Tagen und Monaten erkannt.

Tokenlänge 5 : Das Token wird akzeptiert und zwischengespeichert, wenn es die Form 'x.xx.' oder 'xx.x.' hat. Hat es die Form 'xxxx.' wird der abschließende Punkt entfernt und das Token wird zum Test für Token der Länge 4 weitergereicht. Sonst wird es verworfen.

Hier werden eventuelle Angaben von Tagen und Monaten erkannt. Bei Token der Form 'xxxx.' wird angenommen, dass es sich um Jahreszahlen an einem Satzende handeln kann.

Tokenlänge 4 : Das Token wird akzeptiert und zwischengespeichert, wenn es die Form 'xxxx' oder 'x.x.' hat. Hat es die Form 'xxx.' wird der abschließende Punkt entfernt und das Token wird zum Test für Token der Länge 3 weitergereicht. Sonst wird es verworfen.

Hier werden eventuelle Angaben von Jahreszahlen oder Angaben von Tagen und Monaten erkannt. Bei Token der Form 'xxx.' wird angenommen, dass es sich um Jahreszahlen an einem Satzende handeln kann.

Tokenlänge 3 : Das Token wird akzeptiert und zwischengespeichert, wenn es die Form 'xxx' oder 'xx.' hat. Sonst wird es verworfen.

Hier werden eventuelle Angaben von Jahreszahlen, Angaben von Tagen oder Angaben von Monaten erkannt.

Tokenlänge 2 : Das Token wird akzeptiert und zwischengespeichert, wenn es die Form 'xx' oder 'x.' hat. Sonst wird es verworfen.

Hier werden eventuelle Angaben von Jahreszahlen, Angaben von Tagen oder Angaben von Monaten erkannt.

Token, die nur aus einer einzigen Ziffer bestehen, werden verworfen, da erstens davon ausgegangen wird, dass Tages- und Monatsangaben mit einem Punkt abgeschlossen werden, und zweitens die Wahrscheinlichkeit, dass es sich dabei um eine Jahreszahl, und nicht um irgendeine andere Zahlenangabe handelt, für sehr gering gehalten wird.

Abschließend wird mit Hilfe der Javaklasse SimpleDateFormat und dem DateFormat *SHORT* geprüft, ob das Token eventuell die Form 'tt.mm.jj' [t(ag), m(onat), j(ahr)] hat. Fällt das Ergebnis positiv aus, wird das Token zwischengespeichert, sonst wird es verworfen.

Da es sich bei Ausdrücken der Form 'xx.' um Jahreszahlen am Ende eines Satzes handeln kann, wird der Wertebereich von 'xx' nicht auf 1 bis 31 (maximale Anzahl von Tagen in einem Monat) beschränkt. Die Wahrscheinlichkeit für das Auftreten von Ausdrücken, wie z.B. '99.99.', wird für sehr gering gehalten. Ausdrücke, wie z.B. '9.99', werden schon auf Grund ihres Musters nicht akzeptiert. Dadurch kann es auch nicht zu einer Verwechslung mit gebrochenen Zahlen kommen.

Nach all diesen Tests wird das nächste Token angefordert. Solange ein Token einen der genannten Tests besteht, wird es in den Zwischenspeicher aufgenommen. Sobald ein Token keinen der Tests besteht, wird der Ausdruck im Zwischenspeicher, wenn vorhanden, als Ergebnis der Suche geliefert. Danach wird der Speicher gelöscht und mit den verbleibenden Token fortgefahren. Ein Beispiel wird in Abb. 5 gegeben.

Abb. 5: Beispiel für Extraktion von Datumsangaben

Beispiel : "...am Dienstag, 25. September 2007, wird ..."		
<i>am</i>	→ kein Test bestanden	Speicher : ""
<i>Dienstag</i>	→ in Wortliste gefunden	Speicher : " <i>Dienstag</i> "
25.	→ Test Tokenlänge 3, Form 'xx.' erkannt	Speicher : " <i>Dienstag 25.</i> "
<i>September</i>	→ in Wortliste gefunden	Speicher : " <i>Dienstag 25. September</i> "
2007	→ Test Tokenlänge 4, Form 'xxxx' erkannt	Speicher : " <i>Dienstag 25. September 2007</i> "
<i>wird</i>	→ kein Test bestanden	
<u>Ergebnis</u> : " <i>Dienstag 25. September 2007</i> " Speicher : ""		
...		

4.2.3 Ergebnisse

Zum Testen der Methode wurde das selbe Korpus wie zur Extraktion von Ortsangaben verwendet. Im Testkorpus wurden durch manuelle Suche insgesamt 292 Ausdrücke mit Datumsangaben gefunden. Dabei wurden Intervallangaben als zwei Angaben gezählt. Die Evaluierung der Ergebnisse wurde ebenfalls manuell durchgeführt.

Von den 292 erwarteten Ausdrücken wurden durch die Anwendung des Algorithmus 253 Ausdrücke gefunden. Der Großteil der 39 nicht gefundenen Angaben besteht aus Intervallangaben von Jahreszahlen. Diese wurden auf unterschiedlichste Weise ausgedrückt: xxxx-xxxx , xxxx-xx , xxxx-x , xxx/xxx, xxxx/xx . Andere Angaben, die nicht gefunden wurden hatten folgende Formen: xx./xx. Jahrhundert , xxxer (Jahre) , xx.x , x./x.xx.xxxx .

Der Ausdruck "4th of December 1944" wurde nicht als zusammenhängende Datumsangabe gefunden, da das Wort "of" nicht in der Wortliste steht. Außerdem kamen in den Texten Ausdrücke wie "100 v. Chr." vor, wobei hier nur die Jahreszahl gefunden wurde, und damit nicht die Tatsache erkannt wurde, dass sich diese Datumsangabe auf ein Datum vor dem Jahre 0 bezieht. Für die Aussage "v. Chr." gibt es weitere Varianten wie "vor Chr.", "vor Christi Geburt", "v. Chr. Geb." usw.

Angeht dieses Variantenreichtums von Datumsangaben, wird deutlich, wie komplex die Suche nach ihnen ist. Selbst, wenn in der hier getesteten Methode, nach all diesen Formen gesucht worden wäre, würde der Algorithmus, angewandt auf ein anderes Korpus viele Datumsangaben nicht als solche erkennen können.

Ein anderes Problem stellen die 71 falsch erkannten Ausdrücke dar. Hierbei handelt es sich in den meisten Fällen um Maßangaben wie "10 km" oder "12 Millimeter". Daran wird deutlich, dass auch bei der Extraktion von Datumsangaben der Kontext eines Tokens nicht außer Acht gelassen werden darf. Indem das Vorkommen von Maßangaben hinter Zahlen erkannt würde, könnten solche Fehler vermieden werden.

Ein weiterer Fehler ist im Zusammenhang mit der von Java bereitgestellten Klasse SimpleDateFormat aufgetreten. So wurde zwar ein Ausdruck wie "01.01.2001" richtig geparkt, jedoch wurde auch die Zahl 1000.000.000 als Datumsangabe akzeptiert.

In Tabelle 4 sind die Ergebnisse für die deutschen, englischen und zweisprachigen Texte aufgeführt. Die Einzelergebnisse für jedes Dokument sind im Anhang angegeben.

Tab. 4: Ergebnisse der Extraktion von Datumsangaben

	alle Texte (64)	dt. Texte (36)	engl. Texte (14)	dt./engl. Texte (14)
gesucht	292	155	42	95
true pos.	253	141	32	80
false neg.	39	14	10	15
false pos.	71	51	14	6

In Tabelle 5 wird die Precision und der Recall der implementierten Methode angegeben.

Tab. 5: Precision und Recall bei der Extraktion von Datumsangaben

	alle Texte (64)	dt. Texte (36)	engl. Texte (14)	dt./engl. Texte (14)
Precision	0.78	0.73	0.69	0.93
Recall	0.87	0.91	0.76	0.84

Bei den hier erzielten Ergebnissen kann nicht wirklich eine Unterscheidung von deutsch- und englischsprachigen Texten vorgenommen werden. Die unterschiedlichen Werte für Precision und Recall beruhen auf einer zufälligen Verteilung von Datumsangaben auf die Texte. Die Formen der Datumsangaben sind in allen Texten, in denen sie vorkommen, ähnlich.

5 Abschlussbetrachtung

In dieser Arbeit wurden verschiedene Möglichkeiten zur Extraktion von geografischen und zeitlichen Informationen vorgestellt. Im Anschluss daran wurden zwei einfache, selbstentwickelte Methoden zur Extraktion von Orts- und Zeitangaben beschrieben und ausgewertet. Anhand der Ergebnisse wird ersichtlich, dass diese Methoden noch in vielerlei Hinsicht verbessert werden müssen und können.

Bei der Extraktion von Ortsangaben wurden zwar noch zu viele geografische Referenzen gar nicht erst erkannt, aber das Hauptproblem liegt in der schlechten Precision. Diese entsteht durch die vielen falsch erkannten Ortsnamen (geo-/nongeogeo-Ambiguität).

Bei der Extraktion von Datumsangaben liegt das Hauptproblem in der Vielfalt der Möglichkeiten, ein Datum zu präsentieren.

6 Ausblick

Bei der Extraktion von räumlicher Information sollten in Zukunft deutschsprachige und englischsprachige Texte getrennt voneinander behandelt werden, um besser auf die Eigenheiten der jeweiligen Sprache eingehen zu können. Um die Zahl der nicht erkannten geografischen Referenzen zu verringern, muss den gebeugten Formen des Namens mehr Aufmerksamkeit gewidmet werden, z.B. durch Erkennung des Wortstamms. Bei der Methode zur Extraktion von räumlicher Information wurde im Rahmen dieser Studienarbeit noch nicht auf das Problem der Desambiguierung eingegangen. Anhand der Ergebnisse wird jedoch offensichtlich, dass das unbedingt nötig ist. Wobei diese Ergebnisse nur die Problematik der geo-/nongeogeo-Ambiguität widerspiegeln. Obwohl dieses Problem auch bei englischsprachigen Texten in Bezug auf geografische Entitäten innerhalb Deutschlands auftritt, ist es bei deutschsprachigen Texten sehr viel gravierender. Ein möglicher Lösungsansatz besteht in der Verwendung eines kleineren Wörterbuches, in dem nur Orte ab einer bestimmten Größe aufgeführt sind. Die Gefahr, dass Namen größerer deutscher Städte im normalen Sprachgebrauch etwas anderes als diese Orte benennen, ist sehr gering. Dadurch sinkt natürlich wiederum die Quote der richtig gefundenen Orte. Eine andere

Möglichkeit besteht darin, geografische Namen, die sehr häufig auch im nichtgeografischen Sprachgebrauch vorkommen, zu identifizieren und spezielle Kriterien zu finden, die für oder gegen einen geografischen Gebrauch dieser Wörter sprechen. Da prinzipiell fast alle Namen, die einen geografischen Bezug haben, auch als Namen von Personen auftreten können, ist es unumgänglich ganz allgemeine Kriterien für den Gebrauch geografischer Referenzen zu finden. Dafür ist es nicht unbedingt sinnvoll, starre Listen mit Wörtern zu definieren, die dafür oder dagegen sprechen. Hier eignen sich Methoden des maschinellen Lernens. Um dem Problem der geo-/geo-Ambiguität entgegenzutreten, ist es unbedingt nötig, ein "sauberes" Wörterbuch zu benutzen, in dem jede geografische Entität eindeutig zu identifizieren ist. Das hier benutzte Wörterbuch enthält zum Teil mehrere Einträge zu einer Entität, die sich nur minimal in den Koordinatenangaben oder Bevölkerungszahlen unterscheiden. Um zwei wirklich verschiedene Referenzen mit dem selben Namen voneinander unterscheiden zu können, müssen andere, im Dokument vorkommende, geografische Referenzen dazu in Beziehung gesetzt werden und Kriterien festgelegt werden, nach denen man sich für die eine oder andere Entität entscheiden kann.

Bei der Extraktion von Datumsangaben liegt das Hauptproblem in der Vielfalt der Möglichkeiten, ein Datum zu präsentieren. Um die Anzahl erkennbarer Datumsangaben zu erhöhen, müssen weitere Muster von Datumsformaten in die Suche einbezogen werden. Um andererseits die Zahl der falsch erkannten Datumsangaben zu verringern, muss der Kontext eines Ausdrucks analysiert werden. Da Zahlen, die nicht in einem speziellen Datumsformat, wie z.B. '01.01.08', auftreten, fast immer mit der Angabe einer Maßeinheit versehen sind, und die Anzahl datumsbezogener Wörter bzw. Maßeinheiten beschränkt ist, können Kriterien festgelegt werden, die einen Ausdruck als Datumsangabe auszeichnen oder nicht. So kann zum Beispiel das Auftreten der Maßeinheit "*Millimeter*" als Ausschlusskriterium benutzt werden, aber auch jedes andere aufzählbare Substantiv, das keinen Datumsbezug hat.

7 Literaturliste

- [1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. *Web-a-Where: Geotagging Web Content*. In Proceedings of the 27th SIGIR, July 2004, pp. 273–280.
- [2] Y. Chen, G. Di Fabbriozio, D. Gibbon, R. Jana, S. Jora, B. Renger, B. Wei: *GeoTracker: Geospatial and Temporal RSS Navigation*; Proceedings of the 16th international conference on World Wide Web, May 08-12, 2007, Banff, Alberta, Canada
- [3] D. Fahland, T. M. Gläßer, B. Quilitz, S. Weißleder and U. Leser : *HUODINI – Flexible Information Integration for Disaster Management*, 4th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Delft, NL http://www.informatik.huberlin.de/forschung/gebiete/wbi/research/publications/2007/huodini_final.pdf
- [4] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling, *STEWARD: Architecture of a Spatio-Textual Search Engine*. In ACMGIS 2007, Seattle, WA, Nov. 2007, pp. 186-193.
- [5] I. Mani & G. Wilson. 2000. *Robust Temporal Processing of News*. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000), 69-76. Hong Kong.
- [6] Ch. D. Manning and H. Schütze, *Foundations of statistical natural language processing*, 1999 Massachusetts Institute of Technology; Cambridge, Massachusetts.
- [7] A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, and B. Seeger, *Design and Implementation of a Geographic Search Engine*. In Proceedings of the 8th International Workshop on the Web & Databases, Baltimore, MD, June 2005, pp. 19-24.
- [8] E. Rauch, M. Bukatin, and K. Baker. *A confidence-based framework for disambiguating geographic terms*. In Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, pages 50-54, Edmonton, CA, May 2003.
- [9] F. Schilder & C. Habel. 2001. *From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages*. Workshop on Temporal and Spatial Information Processing at the 39th Annual Meeting of the Association for Computational Linguistics (ACL'2001), 65-72. Toulouse, France.
- [10] <http://www.geonames.org>

8 Anhang

Detaillierte Ergebnisse der Extraktion von Ortsangaben

deutsche Texte

Dok.-Nr.	gesucht	true positives	false negatives	false positives
1	7	6	1	8
3	3	2	1	7
4	7	6	1	6
5	15	11	4	12
7	9	6	3	5
9	2	2	0	0
16	26	17	9	15
20	4	2	2	8
21	5	4	1	7
22	43	31	12	36
23	3	2	1	2
24	8	4	4	13
25	7	1	6	6
30	22	11	11	16
32	8	7	1	14
34	10	7	3	12
35	21	12	9	9
36	0	0	0	15
37	9	4	5	6
38	10	8	2	9
40	5	5	0	7
41	11	8	3	5
42	13	11	2	1
43	2	1	1	1
44	20	15	5	3
48	9	7	2	11
51	4	4	0	29
52	22	9	13	21
53	7	6	1	7
54	29	26	3	6
59	2	1	1	12
60	8	8	0	6
61	11	10	1	9
62	3	1	2	10
63	1	0	1	6
64	1	1	0	7
gesamt	367	256	111	347

englische Texte

Dok.-Nr.	gesucht	true positives	false negatives	false positives
2	15	14	1	0
8	5	5	0	6
10	2	2	0	3
12	12	9	3	7
13	11	8	3	4
14	0	0	0	2
15	2	2	0	0
17	3	3	0	1
18	1	1	0	1
45	14	13	1	2
50	8	7	1	0
55	3	3	0	1
56	2	2	0	0
58	13	13	0	3
gesamt	91	82	9	30

dt./engl. Texte

Dok.-Nr.	gesucht	true positives	false negatives	false positives
6	16	14	2	1
11	0	0	0	5
19	2	2	0	2
26	29	24	5	10
27	27	24	3	4
28	28	23	5	8
29	40	29	11	15
31	31	24	7	45
33	9	9	0	1
39	12	10	2	14
46	4	2	2	2
47	7	5	2	4
49	3	3	0	0
57	3	3	0	1
gesamt	211	172	39	112

Detaillierte Ergebnisse der Extraktion von Datumsangaben

deutsche Texte

Dok.-Nr.	gesucht	true positives	false negatives	false positives
1	0	0	0	0
3	1	1	0	0
4	8	8	0	1
5	7	5	2	2
7	0	0	0	0
9	1	1	0	0
16	15	15	0	2
20	6	4	2	0
21	1	1	0	0
22	18	15	3	9
23	7	7	0	2
24	4	4	0	8
25	5	5	0	0
30	11	11	0	0
32	7	7	0	3
34	3	3	0	0
35	0	0	0	1
36	0	0	0	1
37	3	3	0	0
38	4	2	2	1
40	9	9	0	1
41	4	2	2	2
42	0	0	0	2
43	3	3	0	4
44	4	4	0	4
48	2	2	0	3
51	6	3	3	0
52	14	14	0	0
53	1	1	0	0
54	1	1	0	2
59	3	3	0	0
60	1	1	0	0
61	1	1	0	1
62	5	5	0	2
63	0	0	0	0
64	0	0	0	0
gesamt	155	141	14	51

englische Texte

Dok.-Nr.	gesucht	true positives	false negatives	false positives
2	3	3	0	1
8	4	2	2	0
10	0	0	0	5
12	4	4	0	1
13	13	7	6	0
14	2	0	2	0
15	0	0	0	0
17	4	4	0	0
18	1	1	0	0
45	3	3	0	0
50	1	1	0	0
55	4	4	0	0
56	3	3	0	5
58	0	0	0	2
gesamt	42	32	10	14

dt./engl. Texte

Dok.-Nr.	gesucht	true positives	false negatives	false positives
6	0	0	0	0
11	0	0	0	0
19	0	0	0	2
26	16	10	6	0
27	10	10	0	0
28	14	10	4	0
29	11	9	2	1
31	22	20	2	1
33	3	3	0	0
39	4	4	0	0
46	4	4	0	0
47	6	6	0	0
49	2	2	0	0
57	3	2	1	2
gesamt	95	80	15	6

Selbstständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Studienarbeit selbstständig und nur unter Zuhilfenahme der angegebenen Quellen erstellt habe.

Nora Popp