WIRTSCHAFTSUNIVERSITÄT WIEN
MASTERARBEIT

Titel der Masterarbeit:
Optimierung von Geo-Tagging

Englischer Titel der Masterarbeit:
Optimizing Geo-Tagging

Verfasser:	Philipp Höfer BSc (Wu)
Matrikel-Nr.:	0451611
Studienrichtung:	WINF-M03
Textsprache	Deutsch
Beurteiler:	Univ. Prof. Dipl.-Ing. Mag. Dr. Wolfgang Panny
Betreuender Assistent:	Dipl.-Ing. Mag. Dr. Albert Weichselbraun

Ich versichere:
dass ich die Masterarbeit selbstständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich auch sonst keiner unerlaubten Hilfe bedient habe. dass ich die

Ausarbeitung zu dem obigen Thema bisher weder im In- noch im Ausland (einer Beurteilerin/ einem Beurteiler zur Begutachtung) in irgendeiner Form als Prüfungsarbeit vorgelegt habe. dass diese Arbeit mit der vom Betreuer beurteilten Arbeit übereinstimmt.

Datum

Unterschrift

Abstract

Die vorliegende Arbeit bietet einen neuen Ansatz zur Evaluierung von Geotagging-Ergebnissen unter Berücksichtigung von Benutzerpräferenzen und Beziehungen zwischen "korrekten" und "inkorrekten" Ergebnissen. Dabei wird das Konzept des ökonomischen Nutzens zur Beurteilung des Tagging-Ergebnisses angewandt. Dies gewährleistet eine wesentlich größere und vor allem feiner abgestufte Bandbreite für die Beurteilung. Anstatt das ermittelte Tag lediglich mit korrekt/inkorrekt zu bewerten, kann ein Nutzen im Bereich [0;1] zugewiesen werden. Die Berechnung des Nutzens setzt sich aus zwei Teilen zusammen. Zum einen werden die Tags einer hierarchischen Analyse unterzogen, zum anderen wird via Ontologien nach semantischen Verbindungen zwischen den Tags gesucht. Diese Ontologien können durch benutzerdefinierte Gewichte parametrisiert werden.
Schlussendlich wird ein Test-Framework entwickelt, das diese Überlegungen implementiert und die Möglichkeit Benutzerprofile anzulegen bietet. Dadurch soll gezeigt werden, dass ein Geotagging-Ergebnis für einen Benutzer mehr Nutzen als für einen anderen generieren kann. Das Test-Framework überwacht und dokumentiert die Geotagging-Ergebnisse unter Berücksichtigung der verschiedenen Geotagger-Einstellungen, sodass der Output des Frameworks aufgrund dieser Dokumentation eine konkrete Empfehlung an die entsprechenden Benutzer liefern kann.

Key Words
Geotagging, Geoparsing, Geocoding, Information Retrieval, Information Extraction, test-driven development, Geotagging-Evaluierung

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung

Dieses Kapitel soll einen Überblick über die Arbeit bieten. Zuerst wird ein Einblick zum Thema Geotagging gegeben, der die Notwendigkeit sowie mögliche Schwierigkeiten aufzeigt. Im zweiten Teil wird der Kern der Arbeit sowie die dafür notwendige Vorgehensweise vorgestellt.

1.1 Motivation

Die Vision des Geospatial Web umfasst geographische Daten, die vernetzte Internet-Technologie sowie den sozialen Wandel [Scharl2007]. Projekte wie der Information Diffusion Across Interactive Online Media Media (IDIOM) Media Watch on Climate Change bieten Möglichkeiten an, Web-Seiten sowie einzelnen Artikeln eine zusätzliche Ebene der Referenzierung zuzuweisen – und zwar eine raumbezogene (geospatial). Dadurch eröffnen sich auch viele neue Möglichkeiten für das Data Mining.
Die Verarbeitung räumlich bezogener Daten lässt eine dementsprechende geographische Aufbereitung der Daten zu. Mögliche Anwendungen sind zum Beispiel Suchmaschinen, die Suchabfragen verarbeiten, welche wiederum auf geographisch begrenzte Gebiete beschränkt sind. Korrelationen zwischen Ortsnamen oder auch anderen Bezeichnungen können analysiert werden, um herauszufinden, welche dieser Orte am stärksten mit Begriffen wie Fashion, Essen, Party et cetera in Verbindung gebracht werden. Gerade Location-based Services für mobile Geräte profitieren von diesen Überlegungen.
Dies ist nur ein kleiner Auszug möglicher neuer Applikationen, welche erst durch die Verarbeitung raumbezogener Daten möglich werden. Um dies

umzusetzen ist es nun essentiell aus den zugrundeliegenden Dokumenten die "richtigen" geographischen Referenzierungen zu identifizieren. Spezielle Textmining-Programme, sogenannte Geotagger, versuchen aus unstrukturierten Dokumenten (zum Beispiel Webseiten) geographische Anhaltspunkte zu identifizieren, um diesen Dokumenten einen mehr oder weniger eindeutigen geographischen Fokus zuzuweisen.
Dabei haben die Geotagger mit etlichen Hürden zu kämpfen, wie beispielsweise Ambiguitäten. Ambiguitäten (Zweideutigkeiten) können einerseits vom Typ geo-geo (zum Beispiel Vienna/AT vs. Vienna/US) andererseits vom Typ geo/non-geo sein (Turkey/Staat vs. Turkey/Tier). Diese Beispiele sollen verdeutlichen wie schwer es ist eine 100%-ige Präzision zu erreichen.
Geotagger haben üblicherweise viele Tuning-Parameter, die es möglich machen, Resultsets individuell an die Bedürfnisse anzupassen. Speziell die Größe des vom Geotagger verwendeten Ortslexikon (Gazetteer) hat Einfluss auf die Ergebnisse. Es gilt, die verschiedenen Tuning-Parameter-Änderungen und deren Auswirkung zu dokumentieren beziehungsweise zu überwachen, um auf diesen Erkenntnissen basierend die Entwicklung sowie Anwendung von Geotagger zu verbessern. Schlussendlich soll durch diese Maßnahmen die Präzision der Geotagger erhöht werden.

1.2 Ziel der Arbeit

Die Forschungsfrage lautet: Optimizing Geographic Tagging – Wie kann man das Taggingverhalten, im Speziellen den Algorithmus des Geotaggers optimieren, sodass er für einen bestimmten Anwendungsfall optimale Ergebnisse liefert.
Die Tuning-Parameter, im Speziellen die Gazetteer-Struktur sollen anhand von standardisierten Tests optimiert werden. Daher gilt es im Konkreten ein Test-Framework zu designen, dass die verschiedenen Einstellungen und die daraus erhaltenen Resultate überwacht und dokumentiert.

1.3 Beschreibung der Arbeit

Anfangs wird auf die zugrundeliegende Theorie eingegangen. Hier wird ein Einblick in das Thema gegeben sowie fundamentale Begriffe erklärt. Weiters soll auf die Schwierigkeiten des Taggingverhaltens für geographische Begriffe hingewiesen werden. Schlussendlich werden Forschungsergebnisse aus themenverwandten Arbeiten präsentiert.
Der "praktische Teil" beginnt mit einem Überblick über die zu entwickelnde Architektur sowie verwendeter Technologien (speziell in Verbindung mit der Entwicklung des Test-Frameworks). Danach wird das Design der Testcases (verschiedene Gazetteers), die für die Evaluierung der Geotagger fundamental sind, vorgestellt. Diese Testcases werden weiters in ein Framework eingebunden, das es möglich macht, die verschiedenen Geotagger-Einstellungen zu testen und deren Performance zu vergleichen. Das Framework soll weiters die Möglichkeit bieten, gewisse Benutzerprofile zu erstellen um diesen dann auch spezifische Test-Einstellungen zuordnen zu können.
Zur Entwicklung des Frameworks wird hauptsächlich das *CakePHP*- Framework, sowie *Python* eingesetzt. Die Entwicklung soll auf der Methode des *test-driven-development* basieren. Darunter versteht man die Verwendung agiler Methoden zur Entwicklung. Als agile Methode wird *Extreme Programming* verwendet.

2 Grundlagen

In diesem Abschnitt der Arbeit soll die für die Arbeit wichtige, fundamentale Theorie aufgezeigt werden. Dabei unterteilt sich diese in zwei Teile. Bevor auf die spezifische Extraktion von geographischen Begriffen in Texten eingegangen wird, soll eine kleine Einführung zum Thema "allgemeine" Informationsextraktion gegeben werden.

2.1 Informationsextraktion im Allgemeinen

Die stark anwachsende Anzahl von Texten, insbesondere im Internet, macht es immer schwieriger das darunterliegende, enorme Informationspotential zur Gänze auszufüllen. Welch gewaltiges Wissen sich hier verbirgt, macht eine Studie der School of Information Management and Systems der Universtität Berkeley/Kalifornien aus dem Jahr 2003 deutlich [Lyman2003].
Die Studie zeigt, wieviele Informationen im Jahr 2003 in digitaler Form erfasst wurden. Weltweit waren es fünf Exabyte .
Um der Zahl ein Gesicht zu geben: dies ist ungefähr 37.000 mal der Dokumentenbestand der amerikanischen Library of Congress, welche 17 Millionen Bücher zu ihrem Bestand zählt. 92% der neu generierten Information sind dabei digital. Aufgrund der Tatsache, dass diese Zahlen aus 2003 stammen und seither neue Technologien (durch Web2.0 stark erhöhte Anzahl von "aktiven" Benutzern) aufkamen, kann man erahnen wieviel an Mehr-Information heutzutage im World Wide Web vorhanden ist.
Aufbauend auf der Studie der Universität Berkeley hat die International

Data Corporation[1] (*IDC*) 2008 einen Bericht herausgebracht, wonach das Volumen an digitalen Daten von 180 Exabytes (2006) um das Zehnfache auf 1800 Exabytes im Jahr 2011 anwachsen soll. Weiters wurde erhoben, dass fast 80% aller Daten innerhalb von Unternehmen unstrukturiert beziehungsweise nur teilstrukturiert seien [Biztech2009].
Dies zeigt, welch hoher Bedeutung man dem Thema (automatische) Informationsextraktion (IE) beimessen muss. Die IE beschäftigt sich daher damit, die unstrukturierten Informationen, welche zum Beispiel in hohem Umfang im Internet vorliegen, besser zu erschließen.
Information Retrieval-Systeme geben keine umfassende Analyse des gesamten Inhalts ab, sondern sollen sich nur auf jene Textpassagen konzentrieren, die wesentliche Informationen enthalten. Um dem System mitzuteilen, was relevant ist, muss dies durch vordefinierte domänenspezifische Lexikoneinträge oder entsprechenden Regeln dem System fest vorgegeben werden. [Neumann2001]. Aufgrund einer entsprechenden Konfiguration (zum Beispiel Verwendung eines geographischen Lexikons), kann sich ein Geotagger ausschließlich auf die Extraktion geographischer Informationen konzentrieren. Andere Informationen (zum Beispiel temporale) werden dabei ignoriert.

2.1.1 Kernfunktionalität von IE

Neumann beschreibt die Kernfunktionalität von IE-Systemen anhand zweier Bereiche [Neumann2001]:

- **Eingabe:** Spezifikation des Typs der relevanten Information in Form von Templates (Menge von Attributen) und eine Menge von freien Textdokumenten (Pressemitteilungen, Internet-Dokumente et cetera). Unter Template/Schablone wird das Antwortmuster einer Extraktion, nach den Fragen wer, was, wem, wann, wo und warum verstanden. Dies

[1] IDC ist ein US-amerikanisches Marktforschungs- und Beratungsunternehmen im Bereich Informationstechnologie und Telekommunikation

sind Attribut/Wert-Paare wie zum Beispiel Firmen- und Produktinformationen, Umsatzmeldungen, Personalwechsel, et cetera (siehe Abbildung 2.1).

[*PersonOut PersonIn Position Organisation TimeOut TimeIn*]

Abbildung 2.1: Beispiel eines Templates [Neumann2001]

- **Ausgabe:** eine Menge von instanziierten Templates (Werte für Attribute), die mit den als relevant identifizierten und normalisierten Textfragmenten gefüllt sind. Ein instanziiertes Template ist in Abbildung 2.2 zu sehen.

PersonOut	Dr. Hermann Wirth
PersonIn	Sabine Klinger
Position	Leiter
Organization	Musikhochschule München
TimeOut	heute
TimeIn	

Abbildung 2.2: Beispiel eines instanziierten Templates [Neumann2001]

Im Zuge der MUC[2] (Message Understanding Conference) entstanden folgende Teilbereiche der IE [Pellegrini2006]:

- **Erkennen bekannter Entitäten (Named Entity Recognition, NR):** Ziel ist es ausgezeichnete Entitäten wie zum Beispiel Personenname, Orte, Datum, Geld, Titel aus unstrukturierten Texten zu ermitteln. Als Entität wird ein eindeutig zu bestimmendes Objekt bezeichnet. Entitäten sind somit konkret und eindeutig identifizierbare Ausprägungen eines Entitätstyps (zum Beispiel Entitätstyp = Stadt, Entität = Vienna/Austria/Europe, siehe Abbildung 2.3).

[2]MUC sind Konferenzen, die unter anderem vom US-amerikanischen Verteidigungsministerium initiiert und gefördert werden, mit dem Ziel der Verbesserung der Informationsextraktion.

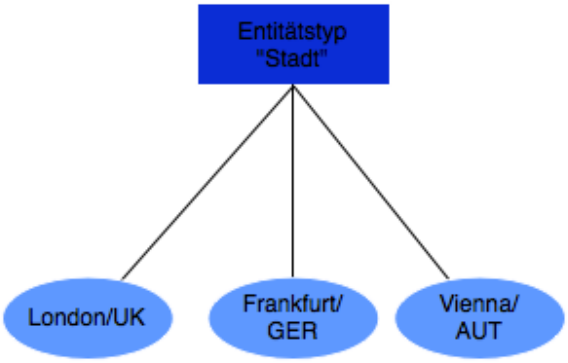


Abbildung 2.3: Entitäten und Entitätstyp

Abbildung 2.4 zeigt diese Überlegungen nun an einem Beispieltext. ”Franz Beckenbauer” stellt hier die textuelle Repräsentation eines Eigennamens (Named Entity) dar und kann der Klasse Person zugewiesen werden. ”München” ist die textuelle Repräsentation eines Eigennamens der Klasse Location.

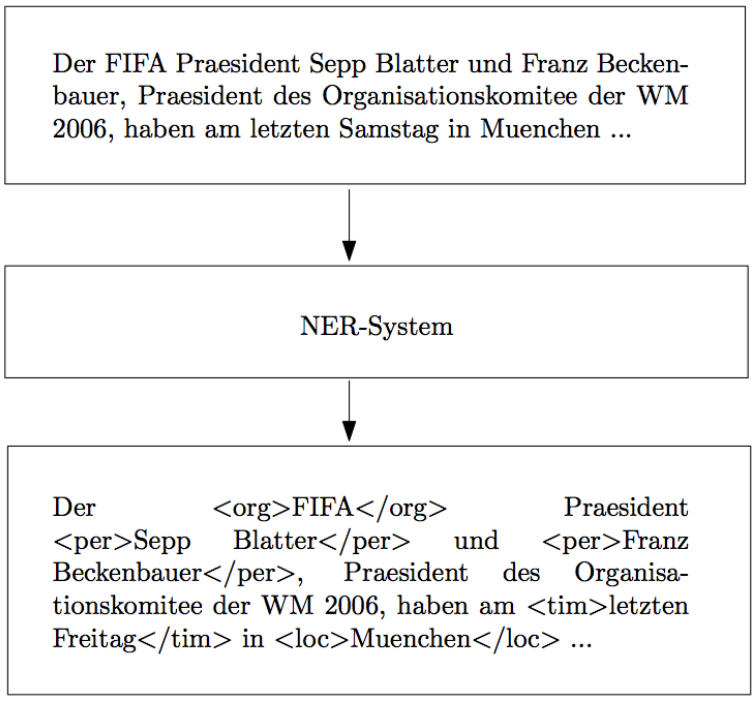


Abbildung 2.4: NER-System: Außenansicht [Lang2006]

Es gibt verschiedene Ansätze zur Eigennamenerkennung (englisch NER - Named Entity Recognition) [Lang2006]:

- **Lexikonbasierter Ansatz:** Es wird ausschließlich ein Wörterbuch verwendet. Über die Wörter hinaus enthält dieses Wörterbuch auch die Zuordnung dieser zu den jeweiligen Klassen. Durch simples Vergleichen mit den entsprechenden Lexikonwörtern werden Wörter oder Wortsequenzen im Text als Eigennamen identifiziert und klassifiziert. Vorteile liegen in der Einfachheit und Schnelligkeit dieses Verfahrens. Schwierigkeiten treten mit Mehrdeutigkeiten (Ambiguitäten) auf. Daher wird dieser Ansatz meist nur in Kombination mit anderen Ansätzen beziehungsweise als Grundlage für weiteres Vorgehen verwendet.
- **Regelbasierter Ansatz:** Wie dem Namen bereits zu entnehmen ist, bilden Regeln und Regelmengen die Grundlage dieses Systems. Regeln bestimmen, wenn eine Sequenz als Eigenname markiert werden soll und zu welcher Klasse ein Eigennamen gehört. Als Regelmengen werden verschiedenste Quellen wie zum Beispiel Lexikonwissen, syntaktisches Wissen, morphologisches Wissen oder weiteres domänenspezifisches Wissen herangezogen. In die Praxis umgelegt, bedeutet dies, dass ein Parser auf Basis der Regeln Eigennamen innerhalb des zu untersuchenden Textes finden soll. Da das manuelle Erstellen dieser Regeln sehr aufwändig ist, wird versucht, diese mittels maschinellem Lernen anzutrainieren. Schwierigkeiten ergeben sich durch die recht spezifische Anwendbarkeit. Regelmengen für Texte in Sprache ”x” können nicht für Texte in Sprache ”y” übernommen werden.
- **Statistische Verfahren:** Anstelle des Regelsystems werden statistische Modelle verwendet. Modelle werden mittels des sogenannten überwachten maschinellen Lernens erzeugt. Im Detail geschieht dies durch (annotierte) Beispiele, welche Schätzungen der zugrundeliegenden Verteilungen erlauben. So können Aussagen für unbekannte Beispiele aufgrund von statistischer Berechnungen ange-

stellt werden. Statistische Verfahren behandeln Wörter meist als atomare Einheiten. Vereinzelt wird aber auch auf Zeichenebene gearbeitet (”Hidden Markov Models”).

- **Auflösen von Co-Referenzen (Co-Reference Resolution, CR):** Die zentrale Aufgabe hier ist es festzustellen, ob unterschiedliche linguistische Objekte auf dieselbe Templateinstanz Bezug nehmen. Folgende Probleme werden in diesem Teilbereich addressiert [Neumann2001]:

  1. EN (Entity-Named)-Koreferenzauflösung stellt beispielsweise fest, dass ”George W. Bush” und ”Bush” in einem Text dieselbe Person bezeichnen.
  2. Promonomiale Referenzen: Referenzen zwischen Pronomen (”er”, ”sie”, ”es”)
  3. Referenzen zwischen Designatoren (”die Firma”, ”der Detroiter Autohersteller”) und anderen Instanzen (”Ford”)

- **Schablonen-Elemente Füllen (Template Element Filling, TE):** Schablonen-Elemente beschreiben zusätzliche Elemente, die zu einer Entität noch extrahiert werden können. So können beispielsweise noch Geburtsdatum und Größe für eine Person extrahiert und zusammen mit dem Personennamen wieder gemeinsam als Schablone zusammengefasst werden.

- **Schablonen-Relationen (Template Relations, TR):** Der Fokus liegt hierbei auf der Extraktion der Relationen zwischen einzelnen Schablonen. So kann diese Relation zum Beispiel zwischen einem Angestellten und dem Bezug zu einer Firma hergestellt werden.

- **Szenario-Schablonen (Scenario Templates, ST):** Ziel ist es hier Szenarien sowie Ereignisse aus freiem Text zu extrahieren, wobei auf die Schablonen-Elemente und Schablonen-Relationen zurückgegriffen wird.

2.1.2 Maschinelles Lernen

Die zuvor angesprochenen maschinellen Lernverfahren lassen sich in mehrere Klassen aufteilen. Man unterscheidet grundsätzlich zwischen **überwachtem Lernen (englisch supervised learning)**, **halb-überwachtem Lernen (englisch semi-supervised learning)** und **unüberwachtem Lernen (englisch unsupervised learning)** [Chapelle2006]. Im Bereich der Eigennamenerkennung kommt zumeist das überwachte Verfahren oder auch halb-überwachte Verfahren zum Einsatz.
Beim automatischen Lernen wird in einem ersten Schritt eine Trainingsbasis erstellt in der eine Reihe von Dokumenten einer Klasse zugeordnet werden. Der Korpus besteht sozusagen aus Instanzen, die den gesuchten Relationen ensprechen und dementsprechend annotiert sind. Neue Dokumente werden aufgrund statistischer Methoden und Lernmodellen den vordefinierten Klassen zugeordnet. Aus einer bereits klassifizierten Teilmenge von Dokumenten werden Merkmale wie zum Beispiel Wörter und Phrasen extrahiert und so dem Klassifizierer als Trainingsbasis zur Verfügung gestellt. Neue Dokumente werden durch den Klassifizierer der Klasse zugeordnet, die am besten mit den Merkmalen des Dokuments übereinstimmt [Glover2002].
Implementierungen dieses Lernverfahrens finden sich in "Hidden Markov"-Modellen, "Maximal Entropy"-Modellen und "Support Vektor Maschinen" (SVM) wieder. Der für Eigennamenerkennung mittlerweile populärste Ansatz ist SVM [Li2005].
Da überwachtes Lernen ausschließlich annotierte Daten betrachtet, muss daher genau in diesem Fall ein anderes Verfahren zur Verwendung kommen. Das halb-überwachte Verfahren addressiert diesen Problembereich. Als Trainingsdaten dient eine Datenbasis aus einer großen Menge nicht-annotierter Daten und einer kleinen Menge annotierter Daten.
Ein Verfahren des halb-überwachten Lernens ist das Self-Training [**?**]. Bei dieser Methode wird zunächst ein Klassifikator aus der kleinen annotierten Menge an Trainingsdaten generiert. Dieser wird weiters dazu verwendet die große Menge an nicht-annotierten Trainingsdaten zu klassifizieren. Nach diesem Vorgang besteht die Menge aus einer Restmenge an den weiter vor-

handen nicht-annotierten Daten sowie aus den (unsicher) klassifizierten Daten. Aus der nun größergewordenen annotierten Menge wird nun wieder ein neues Modell (Klassifikator) gebildet und der gesamte Vorgang wiederholt [**?**].

2.1.3 Evaluationskriterien für IE

Die Güte eines IE-Systems wird anhand zweier Maße beurteilt:

- **Präzision (englisch Precision)**
- **Vollständigkeit (englisch Recall)**

Formel 2.1 zeigt die Definition der Präzision [Lang2006]:

$$P = \frac{Anzahl\ korrekt\ klassifizierte\ NE}{Anzahl\ gefundene\ NE} \tag{2.1}$$

Sie gibt den Prozentsatz der korrekt klassifzierten Named Entities aus der Menge aller gefundenen Named Entities, nicht aber aller vorhandenen Named Entities, an. Sie beschreibt sozusagen die Güte der Klassifikation von Named Entities. Eine hohe Präzision bedeutet daher, dass fast alle gefundenen Named Entities relevant sind [Lang2006].
Die Vollständigkeit ist in Formel 2.2 definiert [Lang2006].

$$R = \frac{Anzahl\ gefundene\ NE}{Anzahl\ vorhandene\ NE} \tag{2.2}$$

Sie misst den Prozentsatz der aufgefundenen Named Entities und damit die Güte der Named Entity Detection-Komponente eines Named Entity Recognition-Systems. Eine hohe Vollständigkeit bedeutet daher, dass fast

alle relevanten Named Entities extrahiert wurden [Lang2006].

Will man nun Präzision sowie Vollständigkeit optimieren, steuert man auf einen Trade-off zu [Neumann2001].
Optimiert man eine Suche auf Präzision hin, so steigt die Wahrscheinlichkeit, dass möglicherweise relevante Wissenseinheiten nicht erkannt werden. Optimiert man andererseits die Vollständigkeit, so steigt die Gefahr, dass Wissenseinheiten mit in das Ergebnis aufgenommen werden, die irrelevant sind. Deshalb wurde ein zusammenfassendes Maß (**F-Maß**) geschaffen, das die Güte eines IR-Prozesses beurteilen soll (siehe Formel 2.3) [Neumann2001]. Es handelt sich dabei um den gewichteten Harmonischen Mittelwert zwischen der Präzision P und der Vollständigkeit R.

$$F = \frac{(\beta^2 + 1) * P * V}{\beta^2 P + V} \tag{2.3}$$

2.2 Extraktion geographischer Information

In diesem Kapitel wird auf die Funktionsweise der Extraktion geographischer Information sowie auf deren Bedeutung eingegangen. Weiters werden insbesonders die dabei auftretenden Probleme und entsprechende Lösungsansätze präsentiert. Abschließend wird die Relevanz dieses Forschungsgebiets anhand alltäglicher Anwendungen aufgezeigt.

2.2.1 Einleitung

Verglichen mit der "klassischen" Information Extraction liegt hier, zusätzlich zu den "klassischen" Anforderungen, der Fokus auf Informationen mit geographischem Kontext. Durch die Miteinbeziehung des räumlichen Aspekts

sollen beispielsweise Suchanfragen Webseiten liefern, die den entsprechenden geographischen Kontext aufweisen. Wenn man betrachtet, dass bereits 2004 20% aller Suchanfragen [Sanderson2004] einen geographischen Kontext aufwiesen, kann man erahnen, welch Potential in diesem Bereich steckt. Die heutzutage starke Verbreitung von mobilen Endgeräten mit der Fähigkeit zum Empfang von Positionssignalen (zum Beispiel GPS) hat einiges dazu beigetragen, dass dieser Anteil noch größer geworden ist. Ein Location-based Service (LBS) wie *Qype Radar*[3] findet nach dem Suchprinzip "Was?" und "Wo?" beispielsweise eine Auswahl von Friseurgeschäften ausgehend von der vom GPS-Empfänger ermittelten Position (siehe Abbildung 2.5). Das Ergebnis kann nach Bewertung oder Distanz sortiert werden.

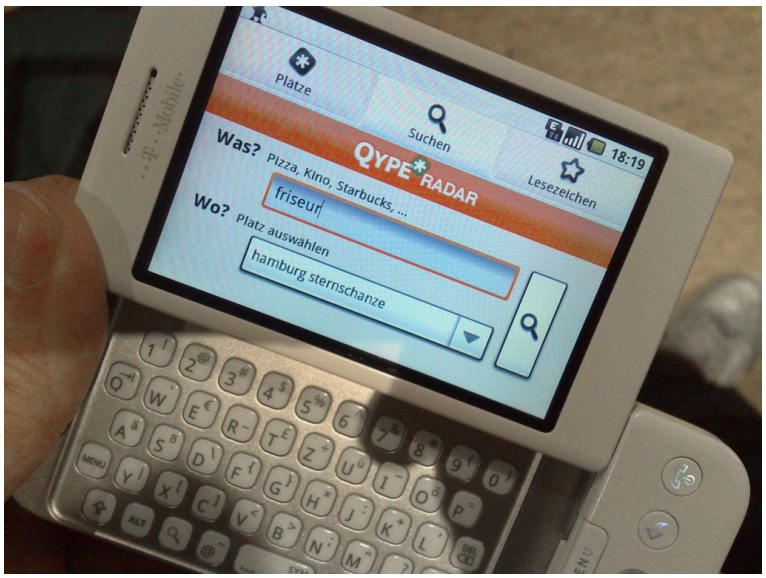


Abbildung 2.5: Abfrage über Friseure in Hamburg [TrendMobi2009]

Herkömmliche Suchdienste, denen bloß textuelle Vergleiche zur Verfügung stehen, können solche Anfragen nur mittels Boolscher Operatoren verarbeiten, zum Beispiel "Pizzeria UND Krems". Die Ergebnisse dieser Suchdienste sind daher oftmals nicht zufriedenstellend. So werden ausschließlich Ergebnisse geliefert, die die Begriffe "Pizzeria" und "Krems" enthalten. Eine mög-

[3] http://mobile.qype.com

licherweise gut bewertete Pizzeria in Stein (Stadtteil von Krems/Donau) könnte so dem Suchergebnis fehlen.
Ein weiteres Problem stellt die Reihung (Relevanzbeurteilung) der Suchergebnisse dar. Diese erfolgt nicht nach geographischen Gesichtspunkten, sondern nach Algorithmen, die auf Zitationsverfahren beruhen. So ist ein Ergebnis (Dokument) umso relevanter, je mehr Webseiten per Hyperlink auf das Dokument referenzieren. Webseiten, welche womöglich den nützlicheren Inhalt für einen Benutzer bereitstellen jedoch georaphisch nur im regionalen Raum auftreten, würden bei diesem Ansatz völlig vernachlässigt werden [Schlieder2006].
Ein anderer Bereich, bei dem der geographische Kontext noch nicht in ausreichendem Maße miteinbezogen wird, ist die Aufbereitung und Verarbeitung von Nachrichten (englisch News). Nachrichtenaggregatoren wie *Yahoo! News*, *Google News* oder *Digg* haben nur rudimentäres Wissen über den geographischen Kontext. Lediglich der Link der Nachricht (Top-Level-Domain wie zum Beispiel ".de", ".at", et cetera) kann zur Bestimmung der Geographie herangezogen werden. Zudem erfolgt die Aufbereitung der Nachrichten zumeist nur aufgrund von Themenbereichen oder Schlüsselwörtern. Ortsbezogene Attribute (Ort, Herkunft, eventuell Wunschdestination des Lesers) des Nachrichtenkonsumenten werden dabei außer Acht gelassen. Teitler et al. addressieren in ihrer Arbeit "Newsstand: A New View On News" dieses Thema und präsentieren eine neuartige, alternative Aufbereitung von Nachrichten nach geographischem Kontext [Teitler2008].
Bei der Umsetzung haben sie zwei Kernfragen identifiziert, welche bei geographisch motivierten Suchanfragen relevant sind [Teitler2008]:

- Feature-based: "Where did story X happen?"
- Location-based: "What is happening in location Y?" (zum Beispiel wichtig für Touristen, die sich über Geschehenes in ihrer Heimat informieren wollen)

Eine weitere große Herausforderung bei der Extraktion stellen Mehrdeutigkeiten dar. Ein Term (zum Beispiel "Paris") kann isoliert betrachtet einen Ort darstellen (Paris/Frankreich) aber auch anders verstanden werden (zum

Beispiel als Vorname einer bekannten Hotelerbin).

Die hier aufgeführten Beispiele sollen die Notwendigkeit der Verarbeitung geographischer Daten nochmals aufzeigen. Damit die aufgeworfenen Fragen beantwortet werden und der darüber hinausgehende geographische Kontext erkannt werden kann, müssen bestehende IE-Systeme um entsprechende Komponenten erweitert werden [Schlieder2006].
Die folgenden Kapiteln adressieren diese Fragestellungen und Anforderungen.

2.2.2 Geographische Entitäten

Bevor auf die Einzelheiten der Extraktion geographischer Information eingegangen wird, sollte der Begriff der "Geographischen Entität" vorab geklärt werden. Speziell für die Behandlung von Ambigutitäten ist es wichtig zu definieren, was als "geo" angesehen wird und was wiederum nicht.
Blotevogel versteht unter Geographie folgendes [Blotevogel2002]:

> ... die Wissenschaft von der Erdoberfläche in ihrer räumlichen Differenzierhung, ihrer physischen Beschaffenheit sowie als Raum und Ort des menschlichen Lebens

Dieser recht breiten Definition gemäß, ist der Interpretationsspielraum für geographische Entitäten sehr groß. Die Abgrenzung geographischer von nichtgeographischer Entitäten ist schwer und kann nicht eindeutig vorgenommen werden. Auch der Scope (Granularität) der Geographie des zu betrachtenden Problems spielt dabei eine Rolle. Im Bereich der Eigennamenerkennung behilft man sich durch Kategorisierungen wie zum Beispiel über Entitätstypen. In [Roth2002] werden die geographischen Entitäten durch Entitätstypen wie Hafen, Flughafen, Insel, Region, Provinz, Land, Kontinent und Gewässer beschrieben.

Die OpenGIS[4] Specification[5] des Open Geospatial Consortium beschreibt ein "geographic feature" als [OGC2009]:

> ...an abstraction of a real world phenomenon; it is a geographic if it is associated with a location relative to the Earth.

Eine ähnliche Definition erfolgt durch Malczewski in [Malczewski1999]:

> A geospatial entity is a term used with respect to an element of a real-world system; that is, entities are contained in geographical space.

Ein einzelnes Objekt ist daher die GIS-Repräsentation einer geographischen Entität. Beschrieben werden diese Objekte durch örtliche und räumliche Daten, sowie nicht-räumlichen Daten, die für die Beschreibung der Charakterisik herangezogen werden.

2.2.3 Quellen geographischer Kontextinformation

Heutzutage werden verschiedenste Verfahren und Ansätze entwickelt, um ein Maximum an geographischen Informationen aus Websites zu extrahieren. Der für die Identifizierung geographischer Kontextinformationen zu einer Koordinate notwendige Prozess wird Geotagging genannt.
Die Quellen dieser geographischen Kontextinformationen können folgendermaßen klassifiziert werden [Scharl2007]:

- **Annotierung durch den Autor:** Diese kann manuell oder auch automatisch erfolgen. Sogenannte Location Aware Devices, wie Navigationsgeräte, Mobiltelefone mit GPS-Module oder auch RFID-fähige Geräte annotieren die erstellte Information automatisch mit geographischen Positionsinformationen.

[4] GIS steht für Geoinformationssystem, also für die Handhabung von räumlichen Daten durch ein rechnergestütztes Informationssystem
[5] Spezifikationen von OpenGIS: `http://www.opengeospatial.org/standards`

- **Identifizierung über den Standort des Servers:** Eine weitere Möglichkeit für die Heranziehung geographischer Quellen stellt die Identifizierung über den Standort des Servers dar. Durch Verwendung des "Whois"-Protokolls[6] können zusätzliche (geographische) Kontextinformationen über eine entsprechende Internet-Domain extrahiert werden.
- **Automatische Annotierung der Seite:** Dieser Prozess splittet sich in zwei Teilprozesse auf. Der erste Teilprozess, genannt **Geoparsing**, versucht aus Volltexten geographische Namen herauszufiltern (zum Beispiel Städte aus einem News-Artikel). Aus den vorhandenen Kontextinformationen werden den identifizierten geographischen Eigennamen die jeweiligen geographische Entitäten zugeordnet. Diese stammen aus einem Ortslexikon, in weiter Folge als Gazetteer bezeichnet. Der zweite Teilprozess umschreibt das **Geocoding**. Hier werden die aus dem Geoparsing ermittelten Referenzen mit Geodaten (Koordinaten) versehen. Damit dies geschehen kann, ist auch hier ein Gazetteer als externe Wissensressource notwendig.
  Natürlich decken Eigennamen (Städte, Länder, Sehenswürdigkeiten, et cetera) einen großen Bereich geographischer Quellen für das Geoparsing ab. Es kann aber darüber hinaus auch aus Anhaltspunkten wie Telefonnummern oder Postadressen Lokalisationsinformationen extrahieren [Ahlers2008].

Bei [Gravano2003] wird beim geographische Kontext zwischen "Location" und "Locality" unterschieden. Location umfasst die geographischen Orte, auf welche in der Webseite Bezug genommen wird. Unter "Locality" versteht man die Beschreibung der geographischen Bedeutung oder auch die Reichweite des geographischen Bezugs. Dieses Konzept hilft zu verstehen und unterscheiden ob der Inhalt einer Webseite für eine lokal und regional begrenzte Nutzergruppe bestimmt ist oder ob dieser ortsunabhänigig und globale Bedeutung verfügt.

[6] Spezifikation: http://www.rfc-archive.org/getrfc.php?rfc=3912

2.2.4 Schwierigkeiten der Geographischen Informationsextraktion

In diesem Kapitel sollen detailliert die Schwierigkeiten des Geographischen Informationsextraktion adressiert werden, wobei speziell auf die Probleme beim Auflösen von Ortsnamen eingegangen wird. Im Folgenden werden verschiedene Problembereiche diskutiert, die Leidner in [Leidner2008] zusammengefasst hat:

Variabilität und Beständigkeit von Ortsnamen

Wie aus der Menschheitsgeschichte bekannt, können Ortsnamen nicht permanent gewissen Territorien oder Grenzen zugeordnet werden. Aufgrund von Kriegen, Abstimmungen et cetera änderten sich die Landesgrenzen des heutigen Österreich über die letzten Jahrhunderte. Umgekehrt kann sich natürlich auch der Name ändern, wobei sich Lage oder Grenzen der geographischen Entität nicht ändert. Durch die Ablösung eines politischen Systems wurde die ostdeutsche Stadt "Chemnitz" in "Karl-Marx-Stadt" getauft und bekam 40 Jahre später wieder ihren ursprünglichen Namen. Dass diese Änderungen auch über einen kürzeren Zeitraum passieren können und somit für große Herausforderungen seitens des Extraktion sorgen, zeigen Städte in Polen. Einige wurden mehr als fünf mal während des zweiten Weltkriegs (sieben Jahre) umbenannt. Weiters kann eine Location auch unter mehreren gebräuchlichen Namen bekannt sein. Orte können in anderen Sprachen anders heißen (Exonyme), Ortsnamen können abgekürzt sein oder in anderen Sprachen sowie auch von der regionalen Bevölkerung anders genannt sein (Endonyme). Der von der lokalen Bevölkerung gebräuchliche Name "Praha" (Endonym) für Prag wird im englischen Sprachgebrauch als "Prague" (Exonym) verstanden.

Ambiguität und Verschwommenheit

Im raumbezogenen Kontext gibt es zumindest drei Typen von Mehrdeu-

tigkeiten:

- **Uneinigkeit:** Diese tritt auf, wenn die Existenz oder Anerkennung von Staaten oder Gebieten in Frage gestellt wird, zum Beispiel Nahostkonflikt Israel-Palästina, Region Kashmir et cetera
- **uneindeutige Spezifikation:** Steht in einem Dokument "A", nördlich von "B" so ist die Location von A noch lange nicht eindeutig. Diese Art der Beschreibung kann dreierlei bedeuten:
  - A befindet sich auf demselben Längengrad wie B, aber eben dem Nordpol näher.
  - A befindet sich nördlich über einem Punkt einer durch B durchgehenden horizontalen Linie.
  - A befindet sich in einem Sektor zwischen Nordost und Nordwest von B.
- **sprachliche Ambiguität:**
  - **Morphosyntaktische Ambiguität:** Ein Wort kann einerseits eine geographische Bedeutung anderseits aber auch andere Bedeutungen haben, zum Beispiel "He is driving to Democrat" versus "She's a democrat". "Democrat' kann einerseits ein Name für einen Ort in North Carolina/USA sein aber eben auch eine politische Gesinnung widerspiegeln. Diese Ambiguität ist vor allem unter dem Namen **geo/nongeo Ambiguität** bekannt.
  - **Merkmalstypen Ambiguität:** Derselbe Ortsname kommt für verschiedene Typen von geographischen Entitäten in Frage. Mit "Wien" kann die Bundeshauptstadt Österreichs aber auch der Fluss gemeint sein.
  - **Referenzielle Amibguität:** Ein Name kann nicht eindeutig einer Location zugeordnet werden: London/England/UK versus London/Ontario/Kanada. Diese Ambiguität findet man unter dem Namen **geo/geo Ambiguität**.

Unter Verschwommenheit ist die intrinsische Unmöglichkeit einer konkreten Zuweisung präziser Grenzen für geographische Entitäten gemeint. Es ist mehr oder weniger unmöglich den Großglockner exakt abzugrenzen.

Metonymie

Darunter wird die Vertauschung von Namen verstanden, sodass ein komplexer Kontext durch ein simpleres Konstrukt erklärt wird. Beispielsweise wird mit der Phrase "Washington sagt, ..." eine Pressemitteilung der amerikanischen Regierung mit Sitz in Washington/DC/USA verstanden. Im raumbezogenen Kontext gibt es drei verschiedene Klassen, unter anderem:

- "place-for-event": Eine Location steht für ein gewisses Event, das dort stattgefunden hat: Waterloo, für Schlacht, die in Waterloo stattgefunden hat.
- "place-for-people": Eine Location steht für eine einzelne oder mehrere Personen, die mit diesem Ort in Verbindung gebracht werden können: London, für Einwohner London's freuen sich über den Zuschlag für die olympischen Sommerspiele
- "place-for-product": Eine Location steht für ein Produkt, welches an dieser produziert wurde: Bordeaux, für den Rotwein der über die Grenzen der Region Bordeaux hinaus weltbekannt ist.

Um Metonymien auflösen zu können ist dementsprechendes Kontextwissen notwendig.

2.2.5 Phasen der Geographischen Informationsextraktion

Die Phasen der Geographischen Informationsextraktion sind in der Literatur weder strikt voneinander getrennt noch genau spezifiziert. In Anlehnung an [Schlieder2006] können aber nach gründlichem Literaturstudium folgende Phasen skizziert werden:

1. Auswertung des Inhalts der Webseiten und die Extraktion von Indikatoren für den geographischen Kontext wie Adressangaben und Ortsbezeichnungen. Diese Phase wird als **Geoparsing** bezeichnet.
2. Weiterverarbeitung der extrahierten Indikatoren und Verknüpfung dieser mit geographischen Koordinaten. Ein weiterer Teil dieser Phase ist die rechnerinterne Repräsentation der geographischen Daten, damit diese für Anwendungen (Visualisierung, Browsing, geographisch fokussierte Suche) genutzt werden können. Diese Phase heißt auch **Geocoding**.
3. **Repräsentation des gesamten geographischen Kontextes** des jeweiligen Dokuments.
4. **Bestimmung des geographischen Fokus (Geofokus)**: Dabei ist die Bestimmung des Raumes gemeint, für den die Informationen des jeweiligen Dokumentes relevant sind.

In der Phase des **Geoparsing** werden aus dem zu analysierenden Dokument Terme extrahiert, die potentielle geographische Konzepte repräsentieren. Geographischer Kontext kann aus Daten wie Telefonnummern oder Adressdaten extrahiert werden [Ahlers2008]. Ahlers beschreibt in seiner Arbeit das Geoparsing von Adressdaten aus Webseiten, um damit eine möglichst präzise Lokalisierung zu erreichen. Diese hohe Granularität (Gebäude-Level) ist besonders für Location-based Service Szenarios nützlich. Problematisch hier ist die Vielzahl an Eventualitäten von Locationinformationen und die mit eingehende hohe Komplexität des zu entwickelnden Geoparsers. Adressinformationen können in verschiedensten Formaten in Dokumenten dargestellt werden. Weiters kann die Adressstruktur über verschiedene Länder hinweg variieren.
Häufigster Ansatz zur Extrahierung geographischen Kontextes ist jedoch die Identifizierung von geographischen Eigennamen. Mittels NER-Methoden werden entsprechende Terme identifiziert.

Die identifizierten Terme werden dann in der Phase des **Geocoding** (auch

Grounding genannt) genau determiniert (Zuordnung von Längen-, Breitengrade, Meereshöhe). Um dies zu bewerkstelligen müssen externe Wissensquellen (Gazetteers) herangezogen werden. Dies sind Thesauri-ähnliche geographische Indizes, die Koordinaten zu entsprechenden geographischen Objekten enthalten [Amitay2004]. Um Adressdaten in einem Gazetteer zu finden wird häufig der Levenshtein-Algorithmus angewendet [Ahlers2008]. Dieser berechnet die Nähe/Distanz zwischen zwei Zeichenketten. Die Anwendung ist notwendig, da Adressdaten häufig unvollständig beziehungsweise abgekürzt vorliegen (zum Beispiel Kärnterstr. statt Kärntnerstraße). Bevor die identifizierten Terme geokodiert werden, werden sie nochmals validiert. Dass hier Schwierigkeiten auftreten können ist leicht ersichtlich. Mehrdeutigkeiten, sogenannte Ambiguitäten, sind in nahezu jedem Dokument enthalten und stellen die größte Problemquelle bei der Informationsextraktion dar.
Ambiguitäten können einerseits vom Typ **geo/geo** als auch vom Typ **geo/non-geo** sein. Geo/geo Ambiguitäten treten auf, wenn derselbe Term mit mehreren geographischen Namen matcht. "Washington" kann einerseits die Hauptstadt der USA sein, andererseits jedoch auch einer der Bundesstaaten der USA sein. Über 25% aller Locations gehören diesem Typus an [Zubizarreta2008]. Geo/non-geo Ambiguitäten sind jene, wenn die Zeichenfolge eines geographischen Ortes auch ein anderes, nicht geographisches Konzept repräsentieren kann. So kann "Paris" in der englischen Sprache die Hauptstadt Frankreichs aber auch der Vorname einer prominenten Hotelerbin sein. Um diese Ambiguitäten zu minimieren gibt es verschiedene Ansätze.
Folgende Filter können angewendet werden um als False-Positives identifizierte Ortsnamen auszuschließen [Zubizarreta2008]:

- Uppercase filter: Diese sollen identifizierte Ortsnamen ausschließen, welche nicht mit einem Großbuchstaben beginnen.
- Stop-Word filter: Ausschließen von häufig vorkommenden Wörtern, die auch geographische Bedeutung haben können. "Oder" wird in einem Dokument durchschnittlich öfters als Konjuktion verstanden als der

Fluss.

- Qualifier filter: Positive und negative Qualifier, die einen Term als Ortsnamen oder nicht identifizieren können. Lokale Präpositionen wie ”an”, ”bei”, ”vor” et cetera sind Indikatoren, dass sich in ihrer unmittelbarer Nähe Informationen geographischen Kontextes befinden können.

Densham und Reid teilen Terme in ”strong terms” und ”weak terms” ein. Als starke Terme werden ausschließlich solche bezeichnet, die auf einen Städtnamen schließen. Schwache Terme repräsentieren nur in Verbindung mit starken ein geographisches Konzept. (zum Beispiel ”Bad” mit ”Vöslau”) [Densham2003] .
Um geo/geo Ambiguitäten aufzulösen gibt es unter anderem kontextuelle Hypothesen wie die des ”single sense of discourse”. Diese Hypothese behauptet, dass ein uneindeutiger Term immer den selben Ort referenziert falls er öfters vorkommt. Kommt in einem Dokument ”Vienna/Österreich” vor, so wird davon ausgegangen dass mit jedem weiterer Auftreten von ”Vienna” ebenfalls die Bundeshaupt Österreichs gemeint ist und nicht die Stadt ”Vienna” im Bundesstaat Georgia der USA. Eine weitere Regel besagt, dass bei einem Auftreten mehrerer mehrdeutiger Orte innerhalb eines gemeinsamen Kontextes diejenigen Orte gemeint sind, die auch räumlich nahe sind. Werden die beiden Orte ”Vienna” und ”Alexandria” innerhalb eines Paragraphs genannt, kann davon ausgegangen werden, dass hier die Städte im Bundesstaat Virginia gemeint sind und nicht ihre bekannteren Namensverwandten in Österreich und Ägypten [Amitay2004].
Ähnlich versucht Leidner in [Leidner2003] das Problem der geo/geo Ambiguität zu lösen. Wie auch beim ”herkömmlichen” Verfahren werden den identifizierten Toponymen (Ortsnamen) die Koordinaten der entsprechenden physischen Orte aus dem Gazetteer zugewiesen. Die identifizierten Toponyme, welche mehreren physischen Orten zugewiesen werden können, werden in einer sogenannten Konfusionsmenge zusammengefasst. Im nächsten Schritt wird das kartesische Produkt aller Konfusionsmengen gebildet, das alle möglichen Kombinationen von eindeutigen Orten enthält. Zum Schluss wird anhand der Koordinaten für jedes Element ein Polygon berechnet. Es

wird angenommen, dass jeweils die Orte in einem Dokument gemeint sind, die das kleinste Polygon bilden.
Abbildung 2.6 zeigt die Idee aus [Leidner2003]. Das aus einem Text identifizierte Toponym A könnte laut Gazetteer den physischen Orten A' oder auch A" zugeordnet werden. Kommen nun in dem Dokument auch die eindeutig zugewiesenen Orte I, J, K vor, so kann angenommen werden, dass der physische Ort A' statt A" in diesem Kontext gemeint ist.

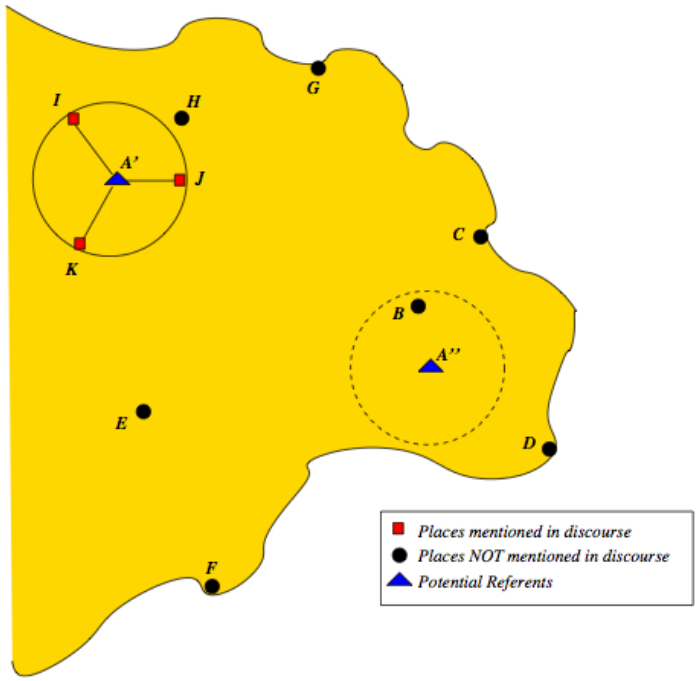


Abbildung 2.6: Geo-geo Disambuigierung nach [Leidner2003]

Nachdem allen Toponymen aus dem Dokument physische Orte zugewiesen worden sind, ist eine geeignete Form der **Repräsentierung des geographischen Kontextes** des Dokuments notwendig. Dies ist wichtig, da die untersuchten Dokumente Bezug auf mehrere Locations nehmen können. Am weitesten verbreitet sind Minimum Bounding Rectangles (MBRs), da sie aufgrund der simplen Struktur einfach und effizient zu handhaben sind. Komplexere geographische Strukturen wie Polygone, konvexe Hüllen, Ellip-

sen et cetera erweitern dieses Konzept [Schlieder2006].
In [Markowetz2005] wird im Kontext einer Suchmaschine von einem geographischen Fußabdruck einer Webseite gesprochen. In diesem sind alle Locations enthalten, die für eine Webseite als relevant erachtet werden. Die Relevanz wird in Form von Wahrscheinlichkeiten für jede Location angegeben. Für die Repräsentation des Fußabdrucks wird die Erdoberfläche mithilfe eines Gitters abgebildet. Dabei wird für jede zu indexierende Webseite die geographische Relevanz für jedes einzelne Feld des Gitters angegeben. Diese Daten werden in einem Index abgelegt und stehen so bei Suchanfragen zur Verfügung.

Die letzte Phase bestimmt den geographischen Fokus oder Gültigkeitsbereich (geographical scope) eines Dokuments oder einer Webseite. Unter dem Gültigkeitsbereich wird der räumlich begrenzte Umkreis, für den die Informationen einer Webseite relevant sind verstanden [Schlieder2006].
Amitay et al. haben in [Amitay2004] einen Algorithmus entwickelt, der es möglich macht, den geographischen Fokus einer Webseite zu bestimmen, wobei dieser gar nicht auf der Webseite vorkommen muss. Die Idee dahinter ist, dass jede aus der Webseite identifizierte geographische Entität einer gewissen Taxonomie unterworfen wird, die für die Steuerung der Granularität von Bedeutung ist. So wird die französische Hauptstadt nicht bloß als physischer Ort "Paris" identifiziert sondern wird in die Form eines taxonomischen Knotens gebracht - Paris/France/Europe. Die einzelnen Hierarchielevel sind verschieden gewichtet, wobei die überstülpenden Hierarchien immer geringer gewichtet sind (France/Europe und Europe). Mittels dieses Ansatzes ist es nun möglich den Geofokus einer Seite, auf der die Orte San Francisco (Kalifornien), Los Angeles (Kalifornien), San Diego (Kalifornien) vorkommen, auf Kalifornien zu setzen. Kalifornien muss dabei gar nicht auf der Seite vorkommen, denn dies wird mittels des beschriebenen Fokusalgorithmus anhand der Taxonomie ermittelt.

2.2.6 Geotagger

Ein Geotagger ist eine Software, die im Prinzip genau die Schritte abhandelt, die in Kapitel "Phasen der Geographischen Informationsextraktion" beschrieben sind. Der Geotagger parst ein Dokument (siehe Abbildung 2.7), extrahiert daraus die Geo-Referenzen und versucht schlussendlich dem Dokument einen geographischen Fokus zuzuweisen.

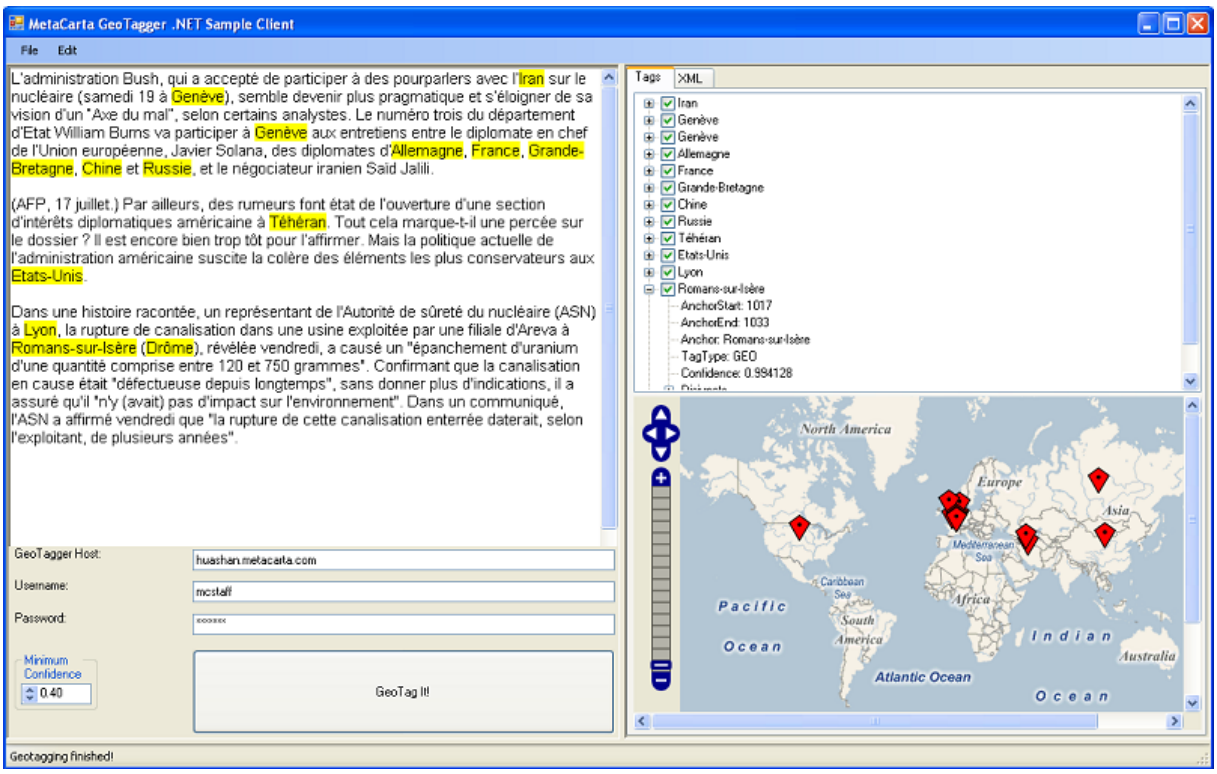


Abbildung 2.7: Beispiel eines Geotaggers [Metacarta2009]

Es gibt aber keine konkreten Standards oder Empfehlungen eines Konsortiums wie die Architektur eines Geotaggers auszusehen hat, daher gibt es verschiedenste Implementierungen. Im Folgenden wird eine mögliche Implementierung eines Geotagger nach [Amitay2004] vorgestellt, der sich auf das Tagging von Ortschaften beschränkt.
Der *Web-a-Where Geotagger* baut dabei auf das *WebFountain* Datamining

Framework [Amitay2004], das von *IBM Research* entwickelt wurde. *WebFountain* stellt einen Webcrawler zur Verfügung, mit dem gefundene Seiten gespeichert und indiziert werden. Zu diesen Seiten können zusätzliche Informationen in Form von Schlüsseln bereitgestellt werden. Die Seiten werden dann mittels Datamining-Tools analysiert. Das Framework unterstützt die Entwicklung von unterschiedlichen Miner. So wurde der *Web-a-Where Geotagger* als eine Variante eines *WebFountain*-Miners, die nach Toponymen sucht, implementiert. Der *Web-a-Where Geotagger* interagiert mit einem Gazetteer, der 75.000 Einträge hat. Die Quellen dieser Einträge sind verschiedene Gazetteers wie *Geographic Names Information System* (*GNIS*)[7], *World Gazetteer* [8], eine *United Nations Division of Sustainable Development* (*UNDSD*)-Länderliste und Ländernamen-Kürzel gemäß ISO-3166-1. Die Auflösung der Toponyme geschieht in vier Schritten. Zuerst werden die extrahierten Begriffe einer Disambiguierung unterworfen. Die geo/non-geo Ambiguität wurde anhand verschiedener Maßnahmen versucht aufzulösen. Der Algorithmus in Listing 2.1 erklärt den Mechanismus. Falls die Begriffe in einem bestimmten Korpus mehr als 100 mal vorkommen und dabei zumeist klein geschrieben sind, wird davon ausgegangen, dass es sich um gewöhnlich englische Wörter und nicht um geographische Entitäten handelt. "Humble" in Texas würde als Term mit einem starken non-geo Bezug gekennzeichnet werden.

```
1 for each entry t in gazetteer do
2   if in corpus, count(t) > 100 && count(uppercase(t)) < count(
      lowercase(t)) then
3     't' might be common English word
4     flag 't' as potential non-toponym
5   end if
6 end for
```

Listing 2.1: Algorithmus für die geo/non-geo Disambiguierung von Amitay et al. [Leidner2008]

[7] enthält mehr als 2 Mio. Einträge der USA
[8] Die erfassten Orte haben in Summe 6,7 Mrd. Einwohner

Eine zweite Regel für die Disambiguierung erfolgt über die Miteinbeziehung von Einwohnerzahlen. Kommt der Name einer möglichen Ortschaft in dem Korpus öfters vor als diese Einwohner hat, so werden diese ignoriert. So kommt es vor, dass "Jordan" als berühmte Persönlichkeit und nicht als Ortschaft auf den Philippinen verstanden wird.
Die geo/geo Disambiguierung und die Vertrauenswerte-Zuweisung wird in Listing 2.2 dargestellt. Unter einem Vertrauenswert wird eine Wahrscheinlichkeit verstanden, mit der ein Name auf einen bestimmten Ort referenziert wird. Im ersten Schritt (Zeile 4-10) werden eindeutig identifizierten Orten (zum Beispiel "Chicago" gefolgt von "IL") Wahrscheinlichkeiten von 95% zugewiesen. Uneindeutige Geo-Entitäten (Springfield, USA) werden im nächsten Schritt behandelt.
In diesem (Zeile 11-14) wird allen unaufgelösten Namen ein Defaultwert, nämlich die Entität mit der größten Bevölkerungszahl und ein Vertrauenswert von 50% zugeordnet.
Der dritte Schritt (Zeile 15-22) steht im Zeichen des "single sense per discourse"-Verfahren, das zuvor schon vorgestellt wurde. Wenn in einem Dokument mehrere Entitäten mit dem selben Namen vorkommen, wovon eine bereits qualifiziert werden konnte, so wird die Bedeutung des qualifizierten Namens an die anderen weiterdelegiert. Man geht davon aus, dass bei einer Mehrfacherwähnung immer derselbe Ort gemeint ist. Gleichzeitig wird ein Vertrauenswert von 90% zugewiesen, wenn dieser Kandidat jener mit der größten Bevölkerungszahl ist - im anderen Fall 80%.
Der letzte Schritt (Zeile 23-32) adressiert die noch übriggebliebenen unaufgelösten Namen - und zwar anhand einer Kontextbetrachtung. Der Kontext wird als Region verstanden, in der die unaufgelösten Namen eindeutig sind. Kommen in einem Dokument die Namen "London" und "Hamilton" vor, so können London/England/UK, London/Ontorio/Kanada, Hamilton/Ohio/USA, Hamilton/Ontario/Kanada und Hamilton/Neuseeland gemeint sein. Der kleinste auflösende Kontext ist bei all diesen Kandidaten Ontario/Kanada. So wird angenommen dass es sich um London/Ontario/Kanada sowie Hamilton/Ontario/Kanada handelt und es werden entsprechende Wahrscheinlichkeiten zwischen 65% und 75% zugewiesen.

```
for each toponym t do
  for each candidate referent tri do
    Initialise confidence scores c(tri) = 0.
    #Local patterns.
    if disambiguating pattern matches then
      #e.g. "Cambridge" and "Cambridge, MA"
      if disambiguation is unique (1 interpretation) then
        Set c(tri) = 0.95 for this candidate referent
      end if
    end if #Maximum population.
    if c(tri) = 0 then
      Assign c(tri) = 0.5 if tri is the referent with the largest
          population.
    end if
    #One-referent-per-discourse.
    if same toponym appears with a disambiguator elsewhere then
      if that referent coincides with the maximum-population referent
          then
        Propagate this referent to all instances with c(tri ) = 0.9
      else
        Propagate this referent with c(tri ) = 0.8
      end if
    end if
    #Geometric minimality.
    #   toponym still unresolved
    find longest common path prefix p in the spatial ontology in which
        all toponyms tu are unambiguous
    for each toponym tu with c(tri) < 0.7 do
      if referent for tu in p coincides with maximum-population
          referent then
        Propagate this referent to all instances with c(tri ) = 0.75
      else
```

```
29          Propagate this referent with c( tri ) = 0.65
30        end if
31      end for
32      Choose referent tri with maximum confidence c(tri).
33    end for
34  end for
```

Listing 2.2: Auflösung von Ortsnamen im *Web-a-Where Geotagger* [Leidner2008]

Der *Web-a-Where Geotagger* wurde auf drei verschiedene Korpora getestet:

- *Arbitrary*: Dieser Korpus besteht aus 200 stark inhaltslastigen Webseiten, welche häufig von Benutzern frequentiert werden. Um den Korpus aufzubauen wurden drei Google-Anfragen (Suchbegriffe: "+the", "+in" und "+and") gestartet und das Ergebnis dieser alphabetisch sortiert. Im nächsten Schritt wurden Webseiten, die eine Größe $<$ 3K hatten, herausgefiltert. Schlussendlich wurden 200 Seiten willkürlich (arbitrary) aus dieser Auswahl selektiert.
- *.GOV*: Aus 1,200.000 Seiten der .gov-Domain wurden 200 Seiten zufällig gewählt. Bei diesen Seiten kann davon ausgegangen werden, dass diese in (grammatikalisch) korrektem Englisch abgefasst wurden.
- *.ODP*: Dieser Korpus besteht aus 200 zufällig gewählten Webseiten aus dem *Regional*-Verzeichnis des Open Directory Project (ODP)[9]. Webseiten mit einer Größe von $<$ 3K wurden herausgefiltert.

Dabei wurden auf insgesamt 600 Webseiten mehr als 7000 Georeferenzen ermittelt. Die Präzisionswerte für die einzelnen Korpora können Tabelle 2.1 entnommen werden. Es wurden vier Testdurchgänge durchgeführt - zuerst wurde der komplette oben beschriebene Algorithmus angewendet, danach wurde jeweils ein Disambiguierungs-Block des Algorithmus deaktiviert um

[9] Das ODP ist das größte von Menschen gepflegte Webverzeichnis des World Wide Web [ODP2010]: http://www.dmoz.org/

Heuristic / Collection	Arbitrary	.GOV	ODP
Full algorithm	81.7%	73.3%	63.1%
No population data	72.7%	68.1%	58.5%
No implied context	82.7%	74.4%	62.0%
No seen-qualified	80.5%	72.3%	61.7%

Tabelle 2.1: Präzision des *Web-a-Where Geotagger* [Amitay2004]

zu beobachten, welchen Einfluss dies auf die Präzision hat. Bei der Anwendung des ganzen Algorithmus konnte eine Präzision zwischen 63% und 82% erzielt werden. Wenn anstatt dem Kandidaten mit der höchsten Bevölkerungszahl ein zufälliger gewählt wird, dann verringert sich die Präzision um bis zu 9% (Maximum-Population-Block). Lässt man die Logik "one-referent-per-discourse"-Block weg so verringert sich bloß die Präzision für den ODP-Korpora um 3,5%. Die anderen beiden erfahren sogar eine Verbesserung. Im Gegensatz dazu steigert diese Maßnahme die Präzision im ODP-Korpora um 3,5%. Lässt man den letzten Block (Geometric Minimality-Block) weg, so fällt bei allen die Präzision, nämlich auf 80,5% bei *Arbitrary*, 72,3% bei *.GOV* sowie 61.7% *ODP*.
Als Fazit dieser Evaluierung kann festgestellt werden, dass die Unterscheidung auf Einwohnerzahl-Ebene im Durchschnitt das größte Potential für die geo/geo Disambiguierung besitzt [Amitay2004].
Die Evaluierungsergebnisse aus [Amitay2004] sind aber mit Vorsicht zu genießen. Anstatt die Ergebnisse gegen einen wiederverwendbaren Gold-Standard zu checken, wird jedes Tag im Nachhinein von einer Person manuell überprüft. Dieses a-posteriori-Vorgehen lässt eine gewisse Subjektivität und damit eine Fehleranfälligkeit zu, welche das Evaluierungsergebnis etwas trübt [Leidner2008].

2.2.7 Gazetteer

Ein Gazetteer ist eine geographische Enzyklopädie, welche für jeden Eintrag einer geographischen Entität zusätzliche Fakten über eben diese speichert.

Dabei können Informationen wie zum Beispiel soziale Statistiken (Arbeitslosenzahlen) oder topologische Informationen (durchschnittliche Seehöhe) über eine Stadt gespeichert werden [Schlieder2006].
Daneben gibt es eine Kurzform dieser Enzyklopädie und zwar den sogenannten short-form gazetteer. Wird in der Literatur von einem Gazetteer gesprochen so ist zumeist dieser *short-form gazetteer* gemeint [Leidner2008]. Ein Eintrag in diesem definiert sich für gewöhnlich durch die drei folgenden Beschreibungen [Leidner2008]:

- **Toponym:** Der Name der geographischen Entität; enthält oft auch verschiedene bekannte Alternativnamen
- **Geographical feature type:** Die Klasse der geographischen Entität; so kann diese ein Land, ein Gebiet, eine Stadt, eine Brücke, ein Flughafen et cetera sein
- **Spatial footprint:** Die Repräsentation der geographischen Entität; die Repräsentation geschieht in den meisten Fällen über die Angabe der Koordinaten von Längen- und Breitengrade. Manche Gazetteere verwenden jedoch auch die Darstellung eines Polygons für die Umrisse der geographischen Entität.

Tabelle 2.2 zeigt eine exemplarische Auswahl von verschiedenen existierenden Gazetteeren.

Alexandria Gazetteer	http://www.alexandria.ucsb.edu/gazetteer
US CIA World Fact Book	https://www.cia.gov/cia/publications/factbook/index.html
Getty Thesaurus of Geographic Names	http://www.getty.edu/research/conducting_research/vocabularies/tgn/
US NGA GEOnet Names	http://geonames.nga.mil/ggmagaz/geonames4.asp
Ordnance Survey (OS) 1:50,000 Scale Gazetteer	http://www.ordnancesurvey.co.uk/oswebsite/products/
State Boundaries of Europe (SBE)	http://www.eurogeographics.org/sbe
United Nations (UNECE) UN-LOCODE	http://www.unece.org/cefact/
US Census Gazetteer	http://www.census.gov/cgi-bin/gazetteer/
Geonames	http://www.geonames.org/

Tabelle 2.2: Liste von Gazetteeren, adaptierte Liste aus [Leidner2008]

Der für diese Arbeit verwendete Gazetteer ist jener von Geonames. Dieser enthält mehr als 8.000.000 geographische Namen welche 6.500.000 einzigartige Topographien entsprechen. Neben dem Ortsnamen sind auch Längengrad, Breitengrad, Seehöhe, Einwohnerzahl, administrative Unterteilung und Postleitzahlen in der Datenbank enthalten. Die Angabe der Koordinaten erfolgt entsprechend dem WGS84-Standard[10], welcher auch zum Beispiel für die GPS-Positionierung verwendet wird. Auf die Daten kann man entweder via Webservices oder über einen täglich aktualisierten Datenbank-Dump zugreifen [Geonames2009].
Der Gazetteer ist einer jener Elemente, die mitunter den größten Einfluss auf das Ergebnis der geographischen Informationsextraktion haben. Leidner beschreibt folgende Kriterien für die Selektion eines Gazetteers, wobei jedes einzeln einen großen Einfluss auf das Ergebnis ausüben kann [Leidner2008]:

- **Gazetteer scope:** Hier wird der Sichtbereich eines Gazetteers festgelegt. Gazetteere können bezüglich ihrer Größe von kleinen Gemeinde-

[10]Spezifikation: http://earth-info.nga.mil/GandG/publications/tr8350.2/wgs84fin.pdf

Datenbanken bis hin zur Abbildung des gesamten Planeten variieren.

- **Gazetteer coverage:** Hier wird zwischen Gazetteeren unterschieden, die eine hohe Dichte an Ortsnamen aufweisen (zum Beispiel NGA GNS), und Gazetteeren die eine eher spärliche Verteilung der Orte innerhalb des Scope aufweisen (zum Beispiel UN-LCODE).
- **Gazetteer correctness:** Dieses Kriterium adressiert Validität und Aktualität der Daten. Wichtig ist, dass die Einträge rasch aufgrund auftretender Ereignisse (zum Beispiel neue Grenzen nach dem Bürgerkrieg im ehemaligen Jugoslawien) angepasst werden.
- **Gazetteer granularity:** Damit wird die Vollständigkeit des Gazetteers beschrieben. Dabei gilt es einen Mittelweg zwischen dem Bias, der durch grob-granulare Gazetteere verursacht wird, und dem Rauschen, welches fein-granulare Gazetteere versuchen, zu finden. Für News-Artikel ist es oft sinnvoller fein-granulare Gazetteers zu verwenden, denn wichtige internationale Ereignisse passieren oftmals in relativ unbekannten Orten (Kampusch-Fall in Strasshof, ca 8.000 Einwohner; Fritzl-Fall in Amstetten, ca. 23.000 Einwohner; Seilbahnunglück in Kaprun -> ca. 3000 Einwohner).
- **Gazetteer balance:** Es ist wichtig einen gleichbleibenden Detailgrad über verschiedene Kontinente oder Regionen zu erzielen.
- **Gazetteer richness of annotation:** Die Detailtiefe eines Gazetteer-Eintrags kann von ausschließlich Koordinaten bis hin zu detaillierten Sozialstatisiken und Einwohnerinformationen reichen.

Abbildung 2.8 zeigt Ambiguitäten innerhalb von Gazetteeren. Die Anzahl der Gazetteer Einträgen ist als Funktion der Anzahl von möglichen Georeferenzen zu sehen. Man sieht, dass es viele Toponyme gibt, die eindeutig zuordenbar sind. Andererseits gibt es auch vereinzeilt Toponyme, die mehr als 1500 (im NGA-Gazetteer) mögliche Referenzen annehmen können. Vergleicht man dieses Ergebnis (1500 Referenzen im NGA-Gazetteer) mit stark grob-granularen Gazetteers wie UN-LOCODE, dann ist bei Verwendung

dieser mit einer möglicherweise inakzeptablen Simplifizierung zu rechnen.

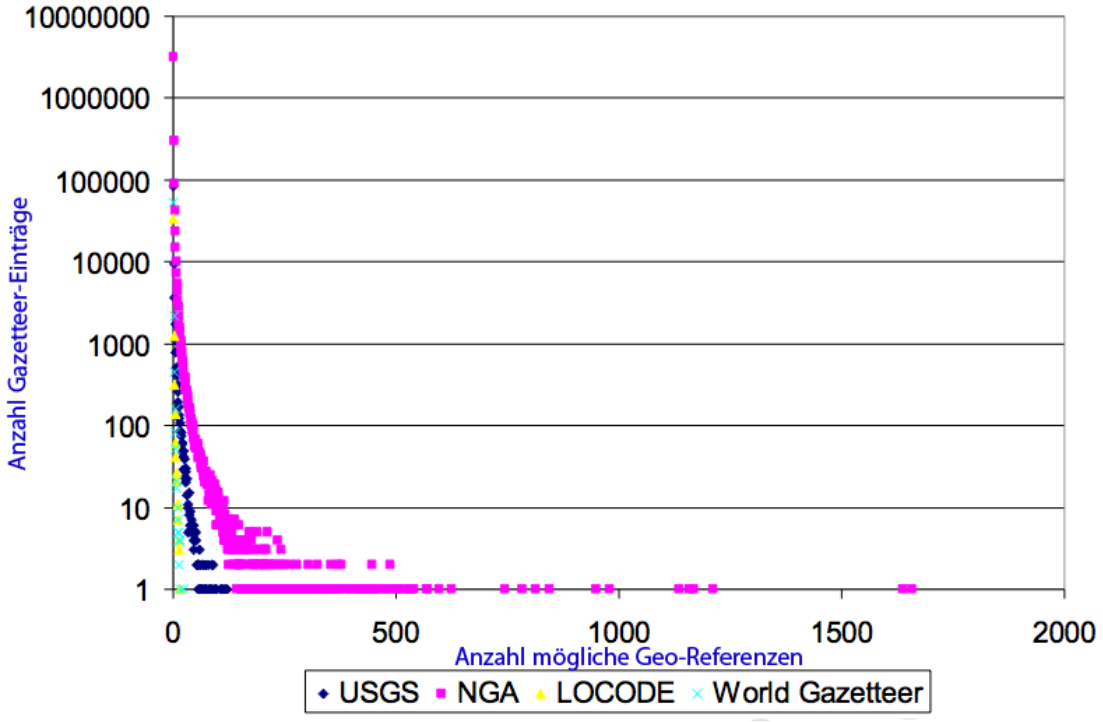


Abbildung 2.8: Gazetteer Ambiguität [Leidner2008]

2.2.8 Beschreibung geographischer Information

Um geographische Daten, besonders im Kontext des Internets, zu speichern oder zu transportieren, benötigt es spezielle Methoden. Die Beschreibung dieser Daten erfolgt durch Auszeichnungssprachen (zum Beispiel **Geography Markup Language**[11], **Keyhole Markup Language**[12]). Die Elemente (zum Beispiel Name der Location, Längengrad, Breitengrad) werden mit speziellen Tags (zum Beispiel <gml:coordinates>45.67, 88.56 </gml:coordinates>) markiert.

[11]Spezifikation: http://portal.opengeospatial.org/files/?artifact_id=20509
[12]Spezifikation: http://portal.opengeospatial.org/files/?artifact_id=27810

Die wichtigsten Beiträge in der Entwicklung von Auszeichnungssprachen für geographische Daten liefert das *Open Geospatial Consortium* (*OGC*)[13]. Diese 1994 gemeinnützig gegründete Organisation hat es sich zum Ziel gemacht, die Entwicklung raumbezogener Informationsdaten auf Basis allgemeingültiger Standards zum Zweck der Interoperabilität voranzutreiben. Das Konsortium besteht aus Mitgliedern von Regierungsorganisationen, Unternehmen und Universitäten. Zu den mehr als 350 Mitgliedern gehören unter anderem *Google*, *Microsoft*, die *Nasa* und *Oracle*.
Im Folgenden wird auf die für die Verarbeitung von Geodaten wichtigsten Standards des *OGC* eingegangen:

Die **Geography Markup Language** (**GML**) ist ein XML-basiertes Meta-Format zur Spezifikation von Austauschformaten für Geodaten. Dies ist der derzeit am breitesten akzeptierte Ansatz zur Beschreibung von Geodaten. Diese Akzeptant drückt sich auch durch die Ernennung zur ISO-Norm 19136 aus, die mit *GML* Version 3.2.1 einhergegangen ist.
Für die Repräsentation der Geodaten wird ein hierarchisch organisiertes Vektordatenmodell verwendet. Dies ist notwendig, da Objekte selbst wieder in anderen Objekten vorkommen können (Länder enthalten Bundesländer, diese wiederum Städte und so weiter). *GML* beinhaltet keinerlei Styleinformation - dies bleibt dem jeweiligen Applikationsdesigner über. Aufgrund des XML-basierten Aufbaus von *GML*, können entsprechende Karten mittels Transformationssprache XSLT erstellt werden. Diese Karten können beispielsweise in Form von skalierbaren Vektorgraphiken (*SVG*) dargestellt werden (siehe Abbildung 2.9) [Behr2009].

[13] http://www.opengeospatial.org/

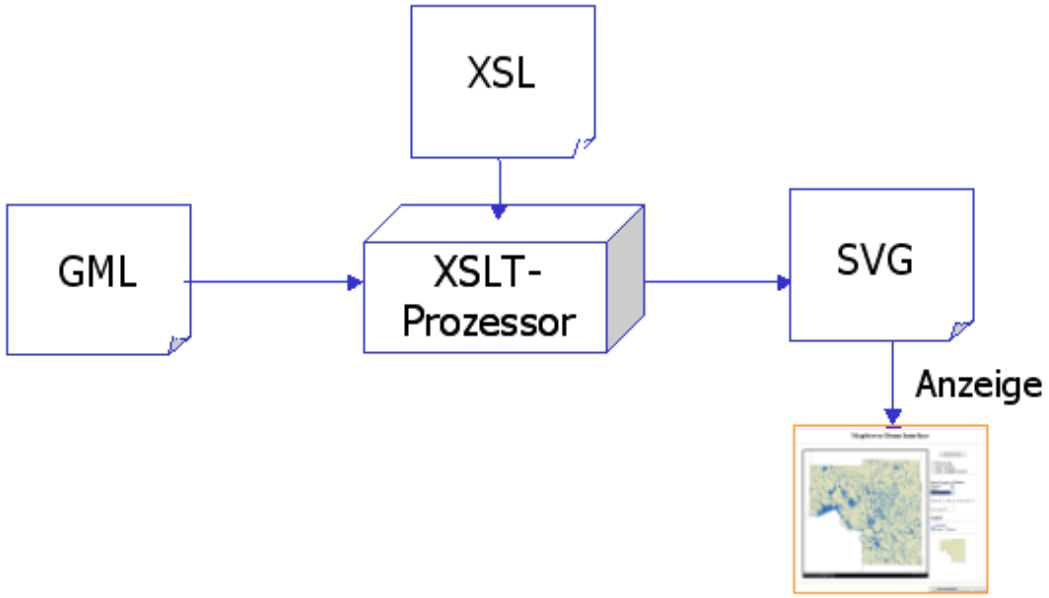


Abbildung 2.9: Transformation von *GML* nach *SVG* [Behr2009]

Die bereits angesprochenen Objekte, in *GML Features* genannt, formen den Kern der Sprache. *Features* bilden die zentrale abstrakte Oberklasse. Davon werden Realweltobjekte, die geometrische- und nichtgeometrische Eigenschaften besitzen, abgeleitet. Die Eigenschaften werden als Properties bezeichnet [Pospech2009]. Unterteilt werden sie in:

- raumbezogene Eigenschaften: diese werden durch Geometrie- und Topologie-Objekte modelliert
- nicht-raumbezogene Eigenschaften: diese können auf zwei Arten modelliert werden: Zum einen können sie durch Attribute mit Standarddatentypen (String, Integer), zum anderen über eine Assoziation von Features zu anderen Klassen modelliert werden.

Eine Straße hat als geometrische Eigenschaft Lage sowie Charakteristik des Straßenverlaufs und als nicht-geometrische Eigenschaft den Namen.
Geometrie-Objekte leiten sich von der abstrakten Oberklasse *Geometry* ab.

Hier können einfache Geometrien wie Punkte, Linien, Polygone, Volumen, zusammengesetzte Geometrien bis hin zu komplexen Interpolationsmethoden (zur realistischen Darstellung von Straßenverläufen) definiert werden. Features können zu einzelnen Einheiten (*FeatureCollections*) zusammengefasst werden [Pospech2009].
So können alle Straßen, Gebäude et cetera auf eine Stadt gemappt werden. Listing 2.3 zeigt ein *GML*-Dokument, indem mehrere Flughäfen zusammengefasst wurden. Die Flughäfen werden durch nicht-geographische Eigenschaften:

(<ogr:NAME>Bigfork Municipial Airport</ogr:NAME>)

sowie durch geographische Eigenschaften beschrieben:

(<ogr:LAT>47.7789</ogr:LAT>)

```
1  <?xml version="1.0" encoding="utf-8" ?>
2  <ogr:Feature Collection
3    xmlns:xsi="http://w3c.org/2001/XML-Schema-instance"
4    xsi:schemaLocation="http://ogr.maptools.org/airports.xsd"
5    xmlns:ogr="http://ogr.maptools.org/"
6    xmlns:gml="http://opengis.net/gml">
7  <gml:boundedBy>
8    <gml:Box>
9      <gml:coord><gml:X>434634</gml:X><gml:Y>5228719</gml:Y>
           </gml:coord>
10     <gml:coord><gml:X>496393</gml:X><gml:Y>5291930</gml:Y>
           </gml:coord>
11   </gml:Box>
12 <gml:boundedBy>
13 <gml:featureMember>
14   <ogr:airports fid="F0">
15     <ogr:geometryProperty><gml:Point>
16       <gml:coordinates>451306,5291930</gml:coordinates>
```

```
17      </gml:Point></ogr:geometryProperty>
18      <ogr:NAME>Bigfork Municipal Airport</ogr:NAME>
19      <ogr:LAT>47.7789</ogr:LAT>
20      <ogr:LON>-93.6500</ogr:LON>
21      <ogr:EVELUATION>1343.0000</ogr:EVALUATION>
22      <ogr:QUADNAME>Effie</ogr:QUADNAME>
23    </ogr:airports>
24  </gml:featureMember>
25  <gml:featureMember>
26    <ogr:airports fid="F1">
27      <ogr:geometryProperty><gml:Point>
28        <gml:coordinates>469137,5271647</gml:coordinates>
29      </gml:Point></ogr:geometryProperty>
30      <ogr:NAME>Bolduc Seaplane Base</ogr:NAME>
31      <ogr:LAT>47.5975</ogr:LAT>
32      <ogr:LON>-93.4106</ogr:LON>
33      <ogr:EVELUATION>1325.0000</ogr:EVALUATION>
34      <ogr:QUADNAME>Balsam Lake</ogr:QUADNAME>
35    </ogr:airports>
36  </gml:featureMember>
37  ....
```

Listing 2.3: Beispiel eines *GML*-Dokuments [Mitchell2008]

Ein weiterer wichtiger Standard, der vom *OGC* spezifiziert wurde, ist *KML* (Keyhole Markup Language)[14]. *KML* ist wie *GML* XML-basiert und hat gerade durch die Verwendung in Applikationen wie *Google Earth* oder *Google Maps* Bekanntheit erlangt. Der Unterschied zu *GML* liegt darin, dass *KML* speziell für die Visualisierung geographischer Daten verwendet wird.
Listing 2.4 stellt einen simplen *KML*-Code zur Beschreibung der Freiheitsstatue in New York dar. Das Dokument ist aus drei Teilen aufgebaut:

[14]Spezifikation: https://portal.opengeospatial.org/files/?artifact_id=27810

- Teil 1: XML-Header
- Teil 2: *KML*-Namespace Deklaration
- Teil 3: Ein Placemark-Objekt, das wiederum aus Attributen wie Name, Adresse oder Point (Koordinaten zur Positionierung auf der Erdoberfläche) et cetera beschrieben wird.

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <kml xmlns="http://www.opengis.net/kml/2.2" xmlns:gx="http://www.
       google.com/kml/ext/2.2" xmlns:kml="http://www.opengis.net/kml
       /2.2" xmlns:atom="http://www.w3.org/2005/Atom">
3  <Placemark>
4    <name>Statue of Liberty</name>
5    <address>National Park Services Liberty Island New York, NY 10004<
         /address>
6    <StyleMap>
7      <Pair>
8        <key>normal</key>
9        <Style>
10         <IconStyle>
11           <heading>0</heading>
12           <Icon>
13             <href>http://maps.gstatic.com/intl/de_ALL/mapfiles/kml/
                   paddle/A.png</href>
14           </Icon>
15           <hotSpot x="0.5" y="0" xunits="fraction" yunits="fraction"/
                 >
16         </IconStyle>
17         <ListStyle>
18           <listItemType>check</listItemType>
19           <ItemIcon>
```

```
20              <state>open closed error fetching0 fetching1  fetching2</
                    state>
21              <href>http://maps.gstatic.com/intl/de_ALL/mapfiles/kml/
                    paddle/A-lv.png</href>
22            </ItemIcon>
23            <bgColor>00ffffff</bgColor>
24            <maxSnippetLines>2</maxSnippetLines>
25          </ListStyle>
26        </Style>
27      </Pair>
28    </StyleMap>
29    <ExtendedData>
30      <Data name="balloon_text">
31        <value></value>
32      </Data>
33      <Data name="contents">
34        <value>Statue of Liberty</value>
35      </Data>
36    </ExtendedData>
37    <Point>
38      <coordinates>-74.044634,40.689062,0</coordinates>
39    </Point>
40  </Placemark>
41  </kml>
```

Listing 2.4: KML-Code Freiheitsstatue New York

Einen Schritt weiter als die bisherigen Methoden zur Beschreibung gehen die Überlegungen rund um das *Semantic Web*. Hier entstehen gänzliche neue Möglichkeiten Metadaten zu beschreiben und zu verwerten. [Schlieder2006]. Mittels einer *Web Ontology Language* (*OWL*) können strukturierte Geodaten mit Nicht-Geodaten zu neuem Wissen verknüpft werden. Die durch *GML* repräsentierten Daten können beispielsweise als In- und Outputparameter der Onotologie eingebunden werden.

Der weltweit bedeutendste offene Gazetteer, Geonames[15], hat eine solche Ontologie entwickelt. In diesem Konzept sind drei Kernklassen zu erwähnen - *Feature*, *Class* und *Code*. *Feature* wird dazu verwendet, um konkrete geographische Entitäten zu beschreiben (USA, London, Eiffelturm, et cetera). Als eindeutig werden sie über einen URI in Geonames identifiziert. Geographische Eigenschaften der Feature Instanzen (Längengrad, Breitengrad, Meereshöhe) werden über die Basic Geo (WGS84 lat/long) Vokabeln der W3C beschrieben (zum Beispiel <wgs84_pos:lat>44.5667</wgs84_pos:lat>). *Class* ist eine Ontologieklasse für die Beschreibung von Konzept-Schemata. Die Geonames Onotology definiert verschiedene Instanzen der Klasse Class - jede repräsentiert ein eigenes Feature Schema. Beispielsweise repräsentiert eine Insanz "H" das Konzeptschema für Seen, Flüsse und andere Gewässer, "T" steht für Berge, Hügel und Gesteine. Die Klasse *Code* bietet ein Set von *Feature Codes* für die verschiedenen Konzept-Schemata. Zur "H"-Instanz der Klasse *Class* gibt es beispielsweise den Code "H.BNK", welcher eine Erhöhung in Gewässer beschreibt (zum Beispiel Sandbänke), die die Navigation schwierig machen.

Das Ziel dieser Disziplin ist es, die räumliche Position der Location auf Webseiten nicht nur anzugeben, sondern durch die Modellierung der Struktur des geographischen Kontextes diesen auf einer höheren semantischen Ebene für Maschinen "verständlich" zu machen [Schlieder2006].
Ein weiteres wichtiges und vor allem einfaches Format zur Auszeichnung von geographischen Informationen auf Webseiten stellt das *geo-Mikroformat* dar. Mikroformate, im Allgemeinen, sind auf Konvention beruhende, HTML-basierte Beschreibungen bestimmter Daten [Microformats2009]. Es gibt mittlerweile verschiedene Mikroformate wie etwa hCard zur Beschreibung von Kontaktdaten oder hCalendar für Termine und kalendarische Informationen. Mittels *properties* werden Längs- und Breitengrade im WGS84-Format definiert (siehe Abbildung 2.10).

[15] http://www.geonames.org/

```
<div class="geo">GEO:
 <span class="latitude">37.386013</span>,
 <span class="longitude">-122.082932</span>
</div>
```

Abbildung 2.10: *geo-Mikroformat* zur Beschreibung der Koordinaten [Microformats2009]

Wie man an diesem Beispiel erkennen kann, liegt der Fokus auf der einfachen Handhabung und Kennzeichnung von Informationen. Mikroformate sind keine neue Sprache, lediglich mit einfacher Semantik angereichertes HTML. So verändert sich auch der Output durch die Miteinbeziehung von Mikroformaten in das HTML-Dokument nicht. Jedoch können Browser-Plugins wie *Operator*[16] Mikroformate automatisch auswerten.
Die prominenteste *geo-Mikroformat* Applikation ist sicherlich das Fotoportal *Flickr*.

2.2.9 Bedeutung von geographischer Information Extraction und Retrieval

Heute (September 2009) kann man das Potential, das in der Verarbeitung von Geodaten beziehungsweise Geotagging steckt, bereits ansatzweise wahrnehmen. Speziell im mobilen Bereich ist in letzter Zeit eine Fülle an Applikationen (beispielsweise das anfangs gezeigte *Qype*), die durch die Verknüpfung von Geodaten mit anderen Daten neuen Nutzen generieren und andeuten was in diesem Bereich alles möglich ist, zu beobachten. Durch Einbeziehung einer weiteren temporalen Dimension sind sogar zeitliche Veränderungen der jeweiligen Daten auf Karten feststellbar.
Wie hoch die Bedeutung beziehungsweise das generelle Potential Geotaggings oder jenes von Geo-aware Applications ist, zeigen nicht nur zuletzt die Prognosen von Gartner[17]. Gartner geht davon aus, dass bis 2012 ein Drittel

[16] http://microformats.org/wiki/operator
[17] Gartner Inc. ist einer der führenden Anbieter für Marktforschung und Analyse in der Technologieindustrie. http://www.gartner.com/technology/home.jsp

aller Analyse-Applikationen für Geschäftsprozesse mit Mashups[18] arbeiten werden [Gartner2009b]. Bei den "Gartner Emerging Trends"[19] wird "Unifying the Digital and Physical Worlds" im Zusammenhang mit Geotagging als eines der großen Themen augelistet [eHomeUpgrade2008].
Abbildung 2.11 zeigt, welche Phasen der öffentlichen Aufmerksamkeit eine Technologie durchläuft [Gartner2009a]. Wie man der Grafik entnehmen kann sind Location-aware Applications in der "Slope of Enlightment" angesiedelt. Dies bedeutet, dass ein gewisser Hype bereits überstanden ist und eine Konsolidierung stattgefunden hat. Vorteile werden weniger emotional betrachtet und man nähert sich der praktischen Nutzbarkeit dieser Technologie. Diese Sichtweise kann leicht nachvollzogen werden, wenn man sich in Erinnerung ruft, dass sich viele Big-Player (*Microsoft*, *Google*) der IT-Welt zusammengefunden haben, um allgemeingültige Standards zur Verarbeitung geographischer Daten zu entwickeln. Weiters haben diese mittlerweile mit *Google Maps* (*Google*) und *Bing Maps* (*Microsoft*) eigene Kartendienste für das Internet herausgebracht.

[18] Als Mashup wird die Erstellung neuer Medieninhalte durch die nahtlose Kombination bereits bestehender Inhalte bezeichnet.
[19] Liste von aufkommenden, wichtigen Technologien für die nächsten fünf bis zehn Jahre

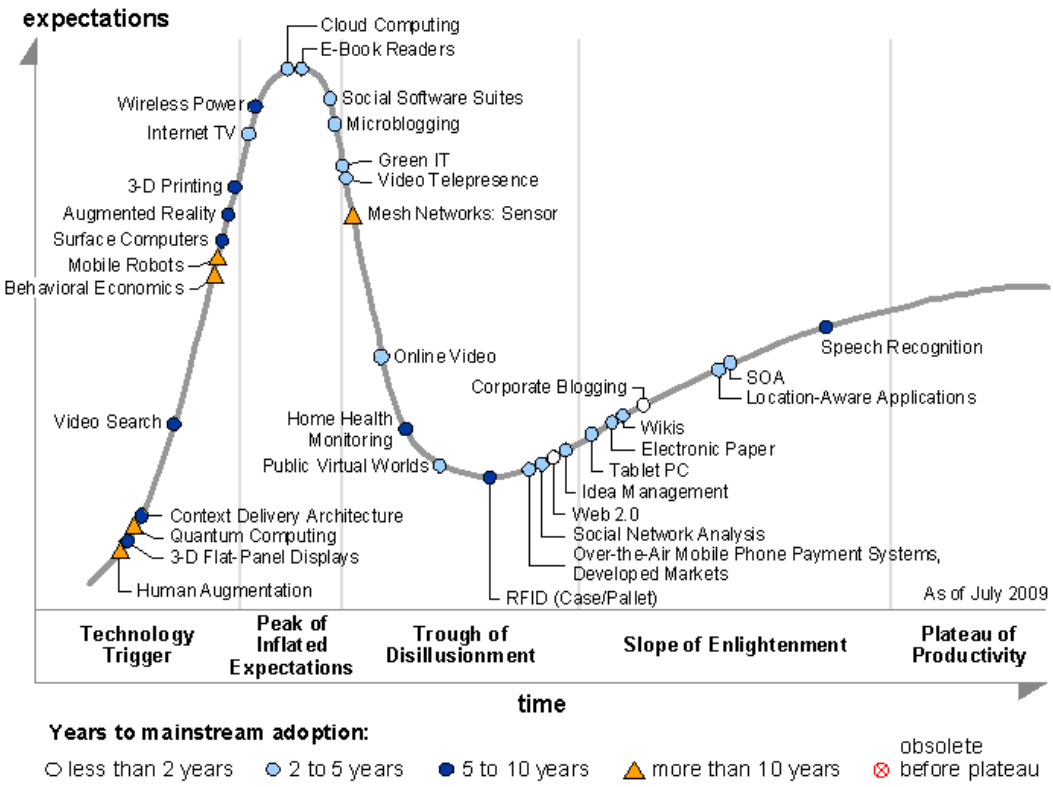


Abbildung 2.11: Gartner Technology Hype Cycle [Gartner2009a]

Dass die Vorteile dieser Technologie mittlerweile praktisch genutzt werden, zeigen die durch Location-based Services/Applikationen generierten Umsätze. Obwohl der Umsatz bei Mobilendgeräten um 4% gegenüber 2008 abgenommen hat, konnte der Umsatz bei Location-based Services mehr als verdoppelt werden ($998,3 Millionen in 2008 auf $2,2 Milliarden in 2009). Als Gründe werden hierfür unter anderem das verstärkte Aufkommen von GPS-Empfängern in Mobilendgeräten sowie die leichte Verarbeitung von Geodaten durch das Ansprechen von APIs (beispielsweise der *Geolocation API* von *Google*) genannt. Gartner sieht für die Zukunft auch einen starken Wachstum im Bezug von Advertising-based LBS. Damit können dem Benutzer Gratis-Dienste angeboten werden, die sich über Werbeeinschaltungen welche wiederum location-based sind, finanzieren. Der Anteil dieser Art von Services soll in den nächsten Jahren von 10-15% (2009) auf 40-50% (2013) anwachsen [Gartner2009b].

Geotagging hat aber nicht nur im kommerziellen Bereich ein hohes Nutzenpotential. Auch Organisationen des Non-Profit Bereichs, Katastrophenmanagements oder des militärischen Bereichs haben dieses Potential zum Teil bereits erkannt. Die US-Army hat erst vor kurzem eine Beta-Version eines Geoparsers[20] entwickelt.
Weiter zukunftsweisende Applikationen könnten sich im Bereich "Augmented Reality"[21] ergeben. Erste Applikationen lassen das große Potential, das sich dahinter verbirgt, erkennen. *Layar*, ein "Augmented Reality"-Browser für das Mobiltelefon-Betriebssystem *Android*, erweitert die Umgebung mit Zusatzinformationen. Über die Kamera seines Mobiltelefons sieht der Benutzer die Bilder seiner Umgebung und die Software markiert bestimmte Punkte oder blendet nützliche Zusatzinformation zu Restaurants, Geschäften, et cetera ein (www.layar.com). Abbildung 2.12 zeigt das Hotel Sacher mit durch *Layar* aufbereiteten Informationen.

[20] GeoDoc: http://geodoc.stottlerhenke.com/geodoc/

[21] Augmented Reality (deutsch: Erweiterte Realität): Darunter wird die computerunterstützte Erweiterung der wahrgenommenen Realität verstanden. Beispielsweise wird während eines Fußballmatches die Entfernung des Freistoßes zum Tor eingeblendet.

Abbildung 2.12: Screenshot von Layar 2.1 [derStandard2009]

Zum Schluss sei noch auf das Potential der neuen Generation von Web-Browsern hingewiesen. Ab Firefox 3.5 lässt sich (bei Zustimmung des Benutzers) die Position des Benutzers ermitteln. Somit können nun auch Benutzer in den Genuss von Location-based Services kommen, ohne auf einen GPS-Empfänger angewiesen zu sein. Dies vergrößert wiederum das Marktpotential dieser Applikationen erheblich.

2.3 Verwandte Arbeiten

In diesem Kapitel sollen einige Arbeiten, die fundamental für die Konzeption dieser Arbeit sind, vorgestellt werden.

2.3.1 Rauch et al.: "A confidence-based framework for disambiguating geographic terms"

Rauch et al. präsentieren in [Rauch2003] einen Ansatz zur Disambiguierung, der auf der Zuweisung von Vertrauenswerten basiert. Dieser Algorithmus findet sich in kommerzieller Anwendung bei Metacarta.
Ortsnamen werden durch die Miteinbeziehung von positiven und negativen Kontexten entsprechend identifiziert und aufgelöst. Dies geschieht per überwachtem Lernen innerhalb eines Trainingskorpus'. Durch diese Vorgehensweise wird jedem Wörterbuchnamen eine sogenannte geographische Signifikanz zugewiesen. Für jedes in Frage kommende Kandidatenpaar (Ortsname n zu Punkt p) wird ein Vertrauenswert (Konfidenz) c berechnet. Dieser gibt an wie hoch die Wahrscheinlichkeit geschätzt wird, dass der Ortsname n mit dem Punkt p übereinstimmt. Jedes dieser p,n-Paare bekommt zuerst eine initiale Wahrscheinlichkeit, die sich aus der durchschnittlichen Wahrscheinlichkeit der Instanz im Trainingskorpus ergibt.
Die Disambiguierung erfolgt aufgrund lokaler und nicht-lokaler Informationen innerhalb des Dokuments. Um festzustellen ob ein Wort zugleich auch eine geographische Entität darstellt, werden lokale Kontextinformationen herangezogen. Starke positive Indikatoren dafür sind Wortgruppen wie "mayor of" oder "community college" mit einem nachfolgenden möglichen Ortsnamen. Negative Indikatoren sind Wörter wie "Mr.", " Dr." oder Vornamen et cetera, wenn diese vor möglichen Ortsnamen stehen. Für die Auflösung der geo/geo Ambiguität (Konfidenz $c_{(p,n)}$) werden nicht-lokale Informationen herangezogen. Beeinflusst wird diese Konfidenz durch die textuelle Nähe zu anderen identifizierten Geoentitäten. Rauch et al. haben herausgefunden, dass zumeist eine Kausalität zwischen der Nähe zweier Orte innerhalb eines Dokuments (zum Beispiel gleicher Absatz) und der räumlichen/geographischen Nähe dieser zwei Orte besteht. Kommen in einem Absatz, die Orte "Seattle" und "Amsterdam" vor, so wird auf Amsterdam im Bundesstaat New York in den USA referenziert anstatt auf die wesentlich bekanntere Hauptstadt der Niederlande. Weiters wirken sich auch die Einwohnerzahlen von p sowie relative geographische Referenzen auf die Wahrscheinlichkeit,

dass n auf einen Punkt p referenziert, aus. Unter diesen Referenzen sind Ortsangaben relativ zu einem Startpunkt zu verstehen (zum Beispiel 15 km nordöstlich von Wien, Österreich).
Diese Erkenntnisse haben Rauch et al. für neue Impulse im Bereich des Information Retrieval genutzt. Dabei werden gefundene Dokumente zusätzlich zur textuellen Relevanz (R_w) auch durch Berücksichtigung einer Georelevanz (R_g) gereiht. Diese ist wie folgt definiert:

$$R_g = C_g \cdot E(P_n, B_n, F_n, S) \tag{2.4}$$

Die Gewichtung E eines Ortsnamens n wird durch seine Position (P_n) im Dokument, seiner Prominenz (B_n), die Frequenz des Namens im Dokument (F_n) sowie die Anzahl der anderen Georeferenzen innerhalb des Dokuments bestimmt (S). Die Prominenz wird durch die Art der Formatierung (fette Schrift, Überschrift, farblich hervorgehoben, et cetera) bestimmt. Ortsnamen, die im Titel oder Header stehen, sind dementsprechend auch höher gewichtet. Die Gewichtungskomponente der Position nimmt von einem Maximalwert zu Beginn eines Dokuments bis kurz vorm Dokumentende ab (Ortsnamen am Schluss werden wieder höher gewichtet). Die Frequenz F_n wird gemäß den Standard-IR-Techniken berechnet. Zum Schluss wird noch die Anzahl anderer Georeferenzen für die Gewichtung eines Ortsnamens miteinbezogen.
Die gesamte Relevanz (R_{total}) eines Terms ergibt sich folgendermaßen:

$$R_{total} = (1 - W_w(|m|))R_g + W_w(|m|)R_w \tag{2.5}$$

$|m|$ ist die Anzahl der Suchterme, W_w ist ein Gewicht um die geographischen Aspekt der Suchanfrage zu konfigurieren, R_g ist die geographische Relevanz

des Terms und R_w ist die textuelle Relevanz, wie sie aus dem "ursprünglichen" IR bekannt ist.

2.3.2 Leidner: "An evaluation dataset for the toponym resolution"

Leidner adressiert die Notwendigkeit von entsprechenden Datenbeständen zur Evaluierung von Geotagging-Algorithmen. In [Leidner2006] wird eine Architektur präsentiert, welche die Erstellung solcher wiederverwendbarer Referenzkorpora unterstützen soll. Auf diese Weise sollen Autoren einfach und benutzerfreundlich sowohl nicht-getaggte als auch bereits nach dem CoNLL-Format getaggte Dokumente geotaggen können.
Abbildung 2.13 zeigt die Architektur, die *Toponym Annotation Markup Editor* (*TAME*) genannt wird. Als Auszeichnungssprache für Locations in Dokumenten wird die XML-basierte Auszeichnungssprache *Toponym Resolution Markup Language* (*TRML*) verwendet.

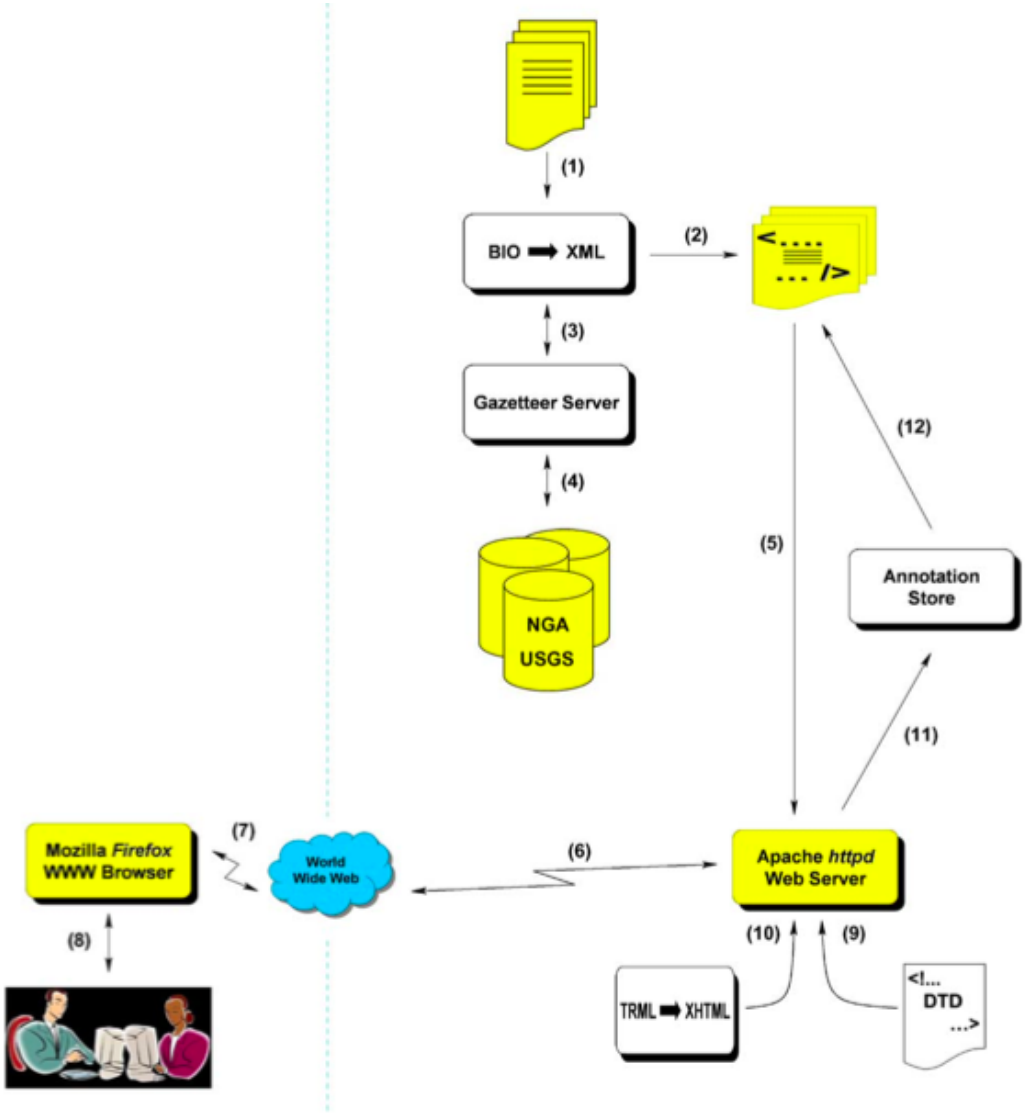


Abbildung 2.13: *TAME*-Architektur [Leidner2006]

Als Datenquelle fungiert ein Auszug (946 Dokumente) aus einem Reuterskorpus, in dem aber mögliche Orte bereits speziell gekennzeichnet sind (CoNLL-Format). Dies hat den Vorteil, dass sich Autoren später nur auf Orte mit auftretender Ambiguität fokussieren müssen. Im ersten Schritt wird das CoNLL-formatierte Dokument mittels eines Perlskripts in *TRML* (siehe Listing 2.5) umgewandelt.

```
<doc id="d1">
  <s id="s1">
    <w tok="EU" pos="NNP" chk="I−NP" ne="I−ORG"/>
    <w tok="rejects" pos="VBZ" chk="I−VP" ne="0"/>
    <w tok="German" pos="JJ" chk="I−NP" ne="I−MISC"/>
    <w tok="call" pos="NN" chk="I−NP" ne="0"/>
    <w tok="to" pos="TO" chk="I−VP" ne="0"/>
    <w tok="boycotts" pos="VB" chk="I−VP" ne="0"/>
    <w tok="British" pos="JJ" chk="I−NP" ne="I−MISC"/>
    <w tok="lamb" pos="NN" chk="I−NP" ne="0"/>
    <w tok="." pos="." chk="0" ne="0"/>
  </s>
  <s id="s2">
    <w tok="Peter" pos="NNP" chk="I−NP" ne="I−PER"/>
    <w tok="Blackburn" pos="NNP" chk="I−NP" ne="I−PER"/>
  </s>
  <s id="s3">
    <toponym did="1" sid="3" tid="1" term="BRUSSELS">
      <w tok="BRUSSELS" pos="NNP" chk="I−NP" ne="I−PER"/>
      <candidates>
        <cand id="c1" src="NGA" lat="−23.3833333" long="29.15"
              humanPath="Brussels &gt; (SF04) &gt; South Africa"/>
        <cand id="c2" src="NGA" lat="−24.25" long="30.95"
              humanPath="Brussels &gt; (SF04) &gt; South Africa"/>
        <cand id="c3" src="NGA" lat="−24.6833333" long="
          26.6833333"
              humanPath="Brussels &gt; (SF04) &gt; South Africa"/>
         [...]
        <cand id="c6" src="NGA" lat="−50.8333333" long="4.3333333
          "
```

```
29                  selected="yes"/>
30            <cand id="c7" src="USGS_PP" lat="-38.94944" long="
                  -90.58861"
31                  humanPath="Brussels &gt; Calhoun &gt; IL &gt; US &gt;
32                      North America"/>
33          </candidates>
34        </toponym>
35        <w tok="1996-08-22" pos="CD" chk="I-NP" ne="0"/>
36      </s>
37      <s id="s4">
38        <w tok="The" pos="DT" chk="I-NP" ne="0"/>
39        <w tok="European" pos="NNP" chk="I-NP" ne="I-ORG"/>
40        <w tok="Commision" pos="NNP" chk="I-NP" ne="I-ORG"/>
41        <w tok="said" pos="VBD" chk="I-VP" ne="0"/>
42      </s>
43       [...]
```

Listing 2.5: *TRML*-Format [Leidner2006]

Jedes <*toponym*> Element beinhaltet eine Kandidatenliste für den identifizierten Term. Um diese Kandidaten zu ermitteln wird in einem entsprechenden Gazetteer nachgeschlagen (Schritt 3,4). Dieses XML-basierte File kann dann über einen Webserver oder einen Fileserver den Autoren zur Verfügung gestellt werden (Schritte 5-8). Da diese jedoch XML spezifisches Wissen dafür benötigen, wird via XSLT (Extensible Stylesheet Language Transformation) das XML-basierte *TRML*-File in HTML umgewandelt (Schritte 9-10). Falls Autoren gegenüber ihrer Tag-Entscheidung unsicher sind, können sie eine entsprechende Checkbox ankreuzen. Abbildung 2.14 zeigt die Maske, die den Autoren zur Verfügung steht.

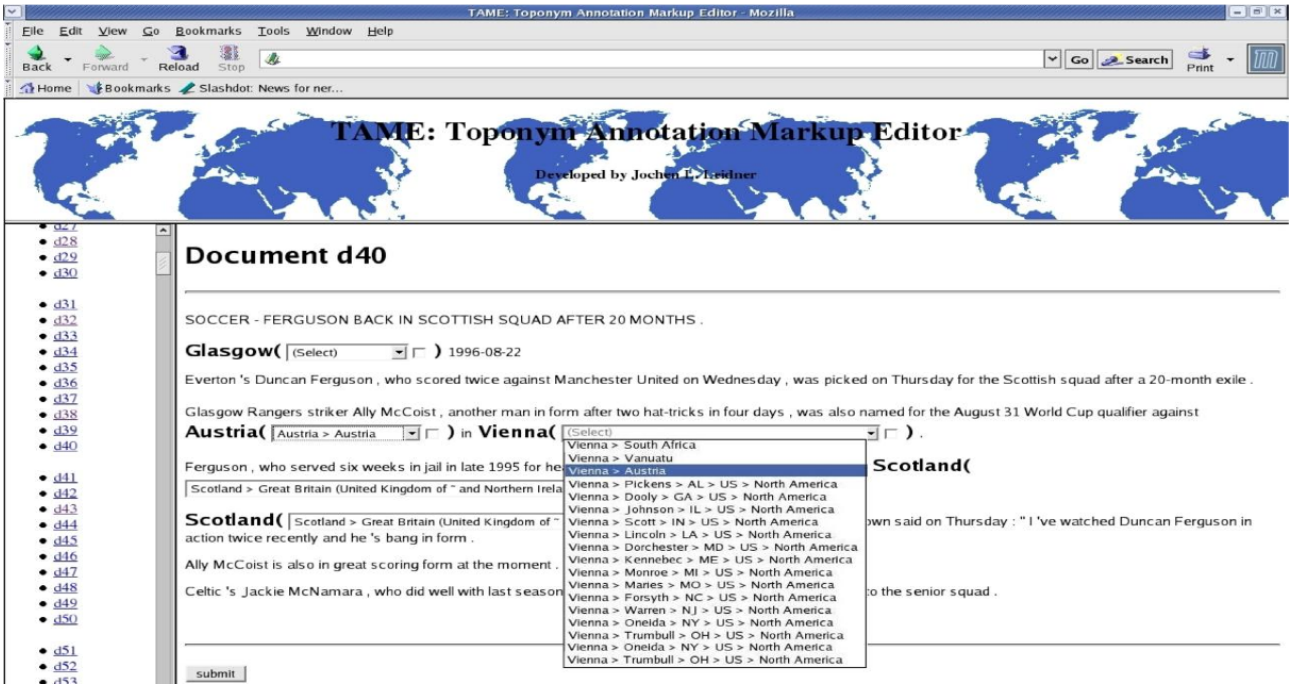


Abbildung 2.14: *TAME*-Editor [Leidner2006]

Nach Abschluss des Taggens eines Dokuments wird dieses mit den Änderungen gespeichert, um danach für eine weitere Überprüfung zur Verfügung zu stehen (Schritt 11-12).

2.3.3 Weichselbraun: "A utility centered approach for evaluating and optimizing geo-tagging"

In [Weichselbraun2009] ist die Basis für die nachfolgende Arbeit enthalten. Weichselbraun präsentiert ein Konzept für ein Test-Framework zur Optimierung des Geotagging-Vorgangs unter Zuhilfenahme der Nutzenfunktion, wie sie aus der ökonomischen Theorie bekannt ist. Damit ist es möglich einem ermittelten Geo-Tag einen fein-granularen Nutzen zwischen null und eins zuzuweisen anstatt wie bei bisherigen Ansätzen bloß zwischen f_{eval}=1

(für korrekt getaggt) und f_{eval}=0 (für nicht-korrekt getaggt) zu unterschieden. Weichselbraun bietet eine neuartige, detaillierte und ganzheitliche Betrachtung zur Bewertung des Nutzens indem er unter anderem Ontologien und benutzerspezifische Einstellungen in die Bewertung miteinfließen lässt. Durch die Miteinbeziehung von spezifischem Wissen (Ontologien) kann die Vergleichbarkeit von Evaluierungs-Metriken für das Geotagging erheblich verbessert werden.
Abbildung 2.15 stellt den Evaluierungsprozess eines Tags innerhalb des Frameworks dar.

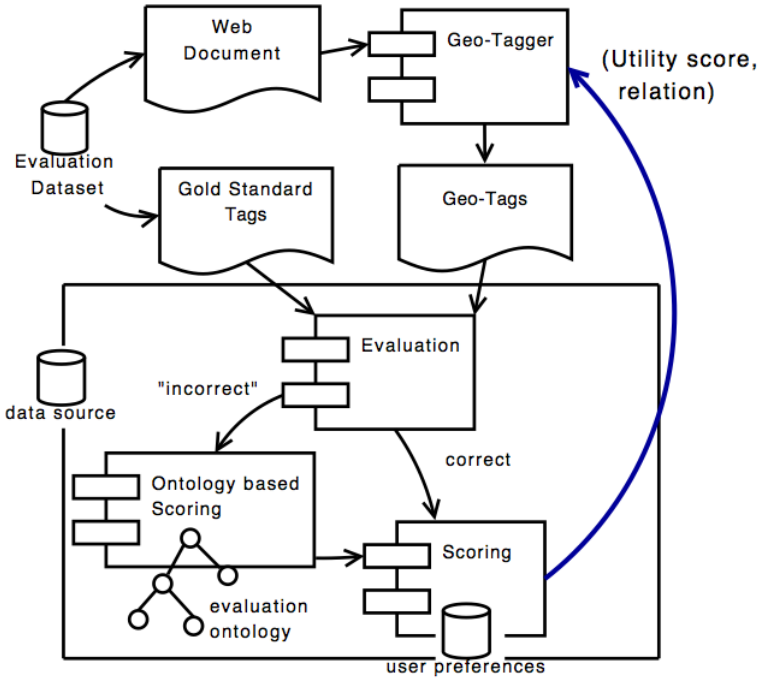


Abbildung 2.15: Aufbau der Evaluierungsarchitektur [Weichselbraun2009]

Die Dokumente eines Evaluation Dataset werden durch den jeweiligen Geotagger getaggt (Fokus wird für Dokument vergeben). Danach werden sie mit einem Gold-Standard Geo-Tag verglichen. Dieses Gold-Standard Geo-Tag entspricht dabei dem richtigen Fokus des Dokuments, da es durch eine (meist) mehr Personen zählende Expertengruppe dementsprechend festgelegt wurde. Korrekt ermittelte Geotags erhalten die volle Punktzahl. Der

Nutzen der nicht-korrekt ermittelten Geotags wird durch zwei Teile bestimmt. Zuerst werden die Geotags einem Ontologie-basierten (siehe Abbildung 2.16) Scoringalgorithmus unter Beachtung der benutzerspezifischen Einstellungen unterzogen um einen Nutzen zwischen Null und Eins zu ermitteln.

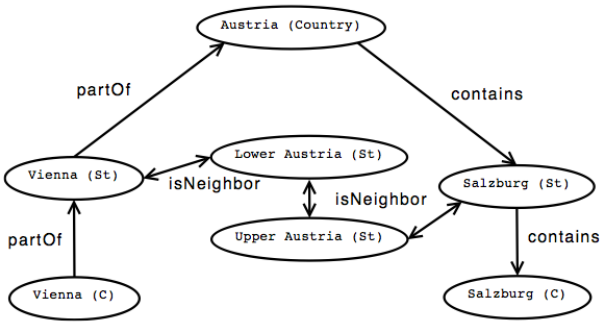


Abbildung 2.16: Ontologie-basierte Evaluierung teilrichtiger Tags [Weichselbraun2009]

Die benutzerspezifischen Einstellungen legen fest wie ”teuer” ein Bewegen entlang der Kanten der Ontologie ist. Der zweite Teil der Nutzenbewertung entspricht der Hierarchieevaluierung. Es werden die übereinstimmenden Hierarchien ermittelt (zum Beispiel ”at” bei ”at/National Park Hohe Tauern” und ”at/Carinthia/Spittal/Heiligenblut”).
Weichselbraun erweitert diesen Ansatz um eine Heuristik, welche auf dem Distanzverhältnis zwischen korrektem und tatsächlichem Tag beruht. Dabei wird das Verhältnis aus zwei Distanzen berechnet. Dividend ist die Distanz zwischen dem korrekten Punkt und dem Punkt aus dem ermittelten Geo-Tag. Als Divisor fungiert die durchschnittliche Distanz zwischen zwei zufällig gewählten Punkten innerhalb einer kreisförmigen Fläche, die jener der zuletzt korrekt ermittelten Hierarchie entspricht (”Austria” bei ”Austria/Upper Austria/Linz” und ”Austria/Styria/Graz”).

3 Umsetzung und Evaluierung

Dieser Teil der Arbeit beschreibt die Umsetzungsphase des entwickelten Test-Frameworks. In den darauffolgenden Kapiteln wird auf Architektur, Algorithmus und entwicklungsspezifische Elemente eingegangen. Zum Schluss werden die Usecases sowie die ermittelten Ergebnisse dargestellt und analysiert.

3.1 Einleitung

Webseiten enthalten oft mehrere geographische Referenzen. Geotagger versuchen anhand der identifizierten Geo-Entitäten einen gewissen Kontext herauszuarbeiten, wobei dieser gleichzeitig auch wieder für die Disambiguierung angewendet wird. Nach dieser Disambiguierungs-Phase versucht ein Geofokus-Algorithmus unter Zuhilfenahme der identifizierten Geo-Entitäten, einem Dokument (Webseite) einen geographischen Fokus zuzuweisen.
Das Ergebnis dieses Fokus-Algorithmus kann anhand verschiedener Parameter beeinflusst werden. So könnte die Veränderung der Minimum-Schranke des Gazetteers von 10.000 (Geo-Entitäten ab einer Einwohnerzahl $\geq$ 10.000 sind im Gazetteer erlaubt) auf 50.000 (Geo-Entitäten ab einer Einwohnerzahl von 50.000) massive Veränderungen mit sich bringen.
Würde eine Webseite die Geo-Entität "Vienna" beinhalten, würde bei ersterer Einstellung (Minimum-Schranke 10.000) und dementsprechendem Kontext möglicherweise Vienna in Virginia/US (Einwohnerzahl ca. 15.000) identifiziert. Bei letzterer Einstellung (Minimum-Schranke 50.000) würde Vienna in Virginia/US gar nicht in Betracht gezogen und so eventuell kein oder

ein möglicherweise unrichtiges Ergebnis geliefert (zum Beispiel Vienna/Austria).

Zusammengefasst kann gesagt werden, dass diese Tuning-Parameter Einfluss auf verschiedene Faktoren wie zum Beispiel die Granularität der Ergebnisse nehmen. So werden bei teilrichtigen Ergebnissen je nach Parameter-Einstellung und Benutzerpräferenz entweder "zu genaue" (zum Beispiel Stadt statt Land) oder "zu ungenaue" (zum Beispiel Land statt Stadt) Ergebnisse geliefert. Diese Ungenauigkeiten machen eine Evaluierung der Geotagger-Performance schwierig.
Ein Artikel über Wolfgang Amadeus Mozart, der die Georeferenzen Salzburg und Vienna (2x) enthält, kann je nach Parametereinstellung folgende Ergebnisse liefern:

- Salzburg: bei "fein-granularer Konfiguration"
- Austria: bei "grob-granularer Konfiguration"
- Vienna: "fein-granulare" Konfiguration" unter Berücksichtigung der Höhe der Einwohnerzahlen

Um dennoch eine möglichst messbare, vergleichbare Größe zur Messung der Geotagger-Performance zu gewinnen, wurde im Rahmen der Entwicklung des Test-Frameworks das **Konzept des ökonomischen Nutzens** der Wirtschaftswissenschaften angewendet [Weichselbraun2009].
Würde man die Performance des Geotaggers nur anhand der richtig getaggten Dokumente beschreiben, so hätte man lediglich die Bewertungsmöglichkeiten null (für unkorrekt getaggtes Dokument) und eins (für korrekt getaggtes Dokument) zur Verfügung.
Ob dabei ein Dokument (Geofokus: "Europe/Austria/Lower Austria") mit "Europe/Austria/Lower Austria/Krems" oder "Asia/India" durch den Geotagger automatisch getaggt wird, würde bei diesem Ansatz bei der Evaluierung des Geotaggers keine Rolle spielen. Beide Tags würden null Punkte lukrieren.

Hätte ein Benutzer gerne Fremdenverkehrsinformationen über Niederösterreich gesucht, würde dieser wahrscheinlich mit einer Webseite über Krems eher zufrieden sein als mit einer Webseite, die Informationen über Indien bietet. Wären in einem durchsuchten Archiv von 1.000 getaggten Webseiten null mit "Europe/Austria/Lower Austria", 150 mit "Europe/Austria/Lower Austria/Krems" und 100 mit "Europe/Austria/Lower Austria/Baden" getaggt, so würde die Suchanfrage des Benutzers leer ausgehen. Dabei beschäftigen sich mehr als 250 Webseiten mit niederösterreichischen Städten, von denen sicherlich auch der Benutzer gerne gewusst hätte.
Gerade die Anwendung der Nutzentheorie kann durch ihre viel breitere Betrachtungsweise (Bewertungspunktsmöglichkeiten [0;1] vs. 0;1) bessere Resultate erzielen.
Eine weitere Beeinflussung des Ergebnisses wird durch **benutzerspezifische Gewichtungen** erreicht. Somit spielt eine weitere Komponente bei der Bewertung des Geotagging-Ergebnisses mit, nämlich die der Benutzerpräferenz. Gemeint ist, dass der Output eines Geotagging-Vorganges bei zwei verschiedenen Benutzern unterschiedlichen Nutzen stiften kann.
Wird ein Dokument (Geofokus: "Europe/Austria/Tyrol/Kitzbühel") mit "Europe/Austria/Salzburg/Obertauern" getaggt, so kann dies für einen Tirol-Liebhaber, welcher ausschließlich Tiroler Schigebiete besucht, ungenügend sein. Andererseits würde sich womöglich eine Person, welche bevorzugt in Salzburg winterurlaubt, das Dokument aufgrund dieses Tags ansehen.

Schlussendlich wird durch die Miteinbeziehung dieser Dimension auch der eigentliche Begriff des Nutzens für den Einzelnen, im Sinne eines Geotagging-Vorganges, auch gerecht.

3.2 Evaluierungsarchitektur

In diesem Kapitel wird zunächst auf die verschiedenen Komponenten der Evaluierungsarchitektur eingegangen. Weiters werden die entwicklungsspezifische Technologien und Bibliotheken vorgestellt. Zum Schluss wird in die-

sem Kapitel die Programmlogik der Evaluierungsarchitektur erläutert, wie etwa sich der Nutzen im Hinblick auf Einflussfaktoren der Benutzerpräferenzen berechnet.

3.2.1 Komponenten

Die Funktionalität des Test-Frameworks wurde in einzelne, eigenständige Komponenten zusammengefasst. Abbildung 3.1 zeigt die Architektur des Test-Frameworks in der Darstellung des UML-Komponentendiagramms.

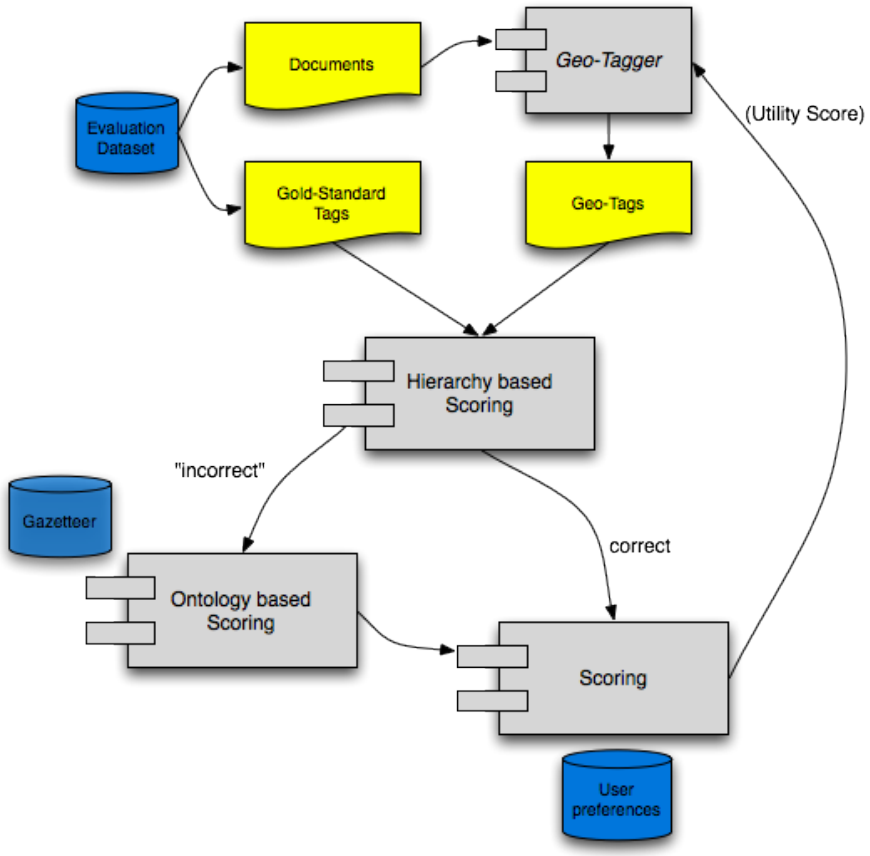


Abbildung 3.1: Architektur des entwickelten Test-Frameworks [vgl. [Weichselbraun2009]]

Der Input des Test-Frameworks (Testcase) besteht aus zwei Elementen:

- **Gold-Standard Geo-Tag**: Dies ist das Tag, das den Fokus für ein Dokument repräsentiert. Die Gold-Standard Geo-Tags wurden mit Hilfe des OpenCalais Geotagger, der via Web-Service angesprochen wird, ermittelt. Der *OpenCalais Geotagger*[1] ist ein nach dem NLP-Verfahren arbeitender Geotagger, der vom Nachrichtendienst Reuters entwickelt wurde.
  Aufgrund mangelnder zeitlicher und personeller Ressourcen einen Korpus manuell zu taggen, entschloss man sich, stützend auf ausgezeichnete Erfahrungswerte, den genannten Geotagger (*OpenCalais*) zum Aufbau eines Referenzkorpusses zu verwenden. Dieser Referenzkorpus besteht aus ca. 15.000 *BBC* Newsartikeln.
- **ermitteltes Geo-Tag**: Dieses Tag wird durch den Geotagger unter Berücksichtigung der bereits genannten verschiedenen Konfigurationen (zum Beispiel Änderung der Gazetteer-Struktur) ermittelt. Als Geotagger wurde hierfür *GeoLyzard* verwendet.

Der Input (Gold-Standard Geo-Tag, ermitteltes Geo-Tag) wird zur Evaluierung an die **"Hierarchy based Scoring"**-Komponente weitergeleitet. Beide Tags unterliegen dabei einem kanonischen Format (Geo-Url). Dieser **Geo-Url** repräsentiert eine Geo-Entität inklusive ihrer Hierarchieebenen. Gemeint ist hierbei die jeweilige Lokation mit all ihren umgebenen Verwaltungseinheiten. Zumindest besteht ein Geo-Url aus einem Kontinent (= höchste Hierarchieebene). Die Stadt Krems in Niederösterreich, Österreich würde in der Geo-Url Schreibweise als Europe/Austria/Lower Austria/Krems dargestellt.
In der Komponente werden beide Tags hingehend ihrer Hierarchien analysiert. Sollte sich das ermittelte Tag als ident erweisen, wird ein Nutzen von eins zu den bisherigen Evaluierungsergebnissen des Evaluation Dataset in der **"Scoring"-Komponente** hinzugezählt.
Im anderen Fall (teilrichtiges Ergebnis) wird das Tag zur weiteren Analyse

[1] http://www.opencalais.com/

in die **"Ontology based Scoring"**-Komponente übergeben. Hier werden die Tags anhand semantischer Zusammenhänge verglichen und daraus dem ermittelten Tag dementsprechender Nutzen zugewiesen. Als Informationsbasis dienen hier externe Quellen in Form von Gazetteers. Zuletzt wird auch das Ergebnis dieser Komponente an die **"Scoring"**-Kompenente zur Berechnung des Gesamtnutzens weitergeleitet.

3.2.2 Entwicklung

Nachdem im letzten Kapitel auf die architektonischen Aspekte des Test-Frameworks eingegangen wurde, werden nun die verschiedenen Technologien, die zum Aufbau jener verwendet wurden, in Form von kompakten theoretischen Exkursen präsentiert:

Entwicklung mit *CakePHP*

Zentrale Bedeutung bei der Entwicklung des Test-Frameworks hat der Einsatz des *PHP*-Frameworks ***CakePHP***.

Folgende Prinzipien bilden den Kern des Frameworks [Scherer2009]:

- ***Rapid Development***: Das Entwickeln von Applikationen besteht aus mehreren Phasen, wovon die reine Programmierung nur eine dieser ist. Weitere wichtige Phasen neben der Implementierung stellen Analyse der Anforderungen, Anwendungsfälle und Testen dar. Der Entwicklungsprozess ist damit einer linearen Ordnung unterworfen und wird in der Regel chronologisch durchlaufen. Diese lineare Vorgehensweise hat neben einigen Vorteilen (gute Möglichkeiten zur Planung und Kontrolle) klare Nachteile. Durch die starre nicht-iterative Struktur ist dieses Modell, gerade wenn man Kunden bedienen will, wenig geeignet. Der Kunde würde das Produkt erst nach der Testphase, also wenn

es fertig ist, präsentiert bekommen. Änderungswünsche können so nur zeit- und kostenintensiv eingearbeitet werden.
Im Gegensatz dazu basiert das Prinzip des *Rapid Development* auf der Idee, die Planungsphase bei der Entwicklung einer Software kurz zu halten um rasch Code für die Erstellung von Prototypen zu generieren.
Das Ziel ist es, dem Kunden so früh wie möglich ein funktionsfähiges Modell zeigen zu können, um etwaige Missverständnisse frühestmöglich auszuräumen.
In *CakePHP* wird durch sogenanntes Prototyping genau dieser Philosophie Rechnung getragen. So kann man nach Definieren eines Datenmodells (zum Beispiel Datenbanktabelle) mit wenigen Befehlen eine webbasierte Oberfläche mit den elementaren Datenbankoperationen (Erstellen, Anzeigen, Editieren, Löschen) erstellen, ohne dabei auch nur einen einzigen SQL-Befehl zu definieren. In *CakePHP* wird dies *CRUD* (Create, Read, Update, Delete) genannt.

- ***"DRY - Don't Repeat Yourself"***: Hier geht es um die Vermeidung sinnloser Redundanzen im Quelltext beziehungsweise in der Datenhaltung. Das Ziel ist es schlanke Applikationen zu schaffen, die einfach zu warten sind, ohne das gleiche oder ähnliche logische Bereiche nebeneinander existieren. Somit soll die Bildung von Seiteneffekten bei der Wartung der Software vermieden werden.
  CakePHP unterstützt dies indem gewisse Standardaktionen (Authentifizierung, Datenzugriff, Session-Verwaltung) schon im Framework abgewickelt werden. Der Programmierer kann sich so auf die eigentliche Applikationenlogik konzentrieren und muss das Rad (zum Beispiel Authentifizierung) nicht neu erfinden.
- ***"Convention over Configuration"***: Der Entwickler muss nur die Bereiche einer Applikation konfigurieren, die sich nicht aus dem logischen Zusammenhang ergeben. Gibt es eine Datenbanktabelle namens *users*, dann wird der zuständige Controller unter *users_controller.php* abrufbar sein und greift auf Templates innerhalb des Ordners *users*

zu. Wenn der Entwickler sich an diese Namenskonventionen hält, muss stattdessen dafür kein spezieller Code geschrieben werden.

Mit der Verwendung des *CakePHP*-Frameworks wurde versucht, das mit-einhergehende Design Pattern des **Model-Viewer-Controller (MVC)** so weit wie möglich in die Applikation zu integrieren.
Das Ziel dieses Patterns ist es, eine logische Trennung der Funktionalitäten in einer interaktiven Applikation in abgegrenzte Teilsysteme herbeizuführen. Die Verwendung dieses MVC-Patterns findet sich heute in verschiedensten Frameworks wieder, zum Beispiel im populären *Ruby on Rails*[2]. Durch diese Trennung kann eine hohe Flexibilität sowie Wiederverwendbarkeit erzielt werden. Abbildung 3.2 zeigt dieses Modell.

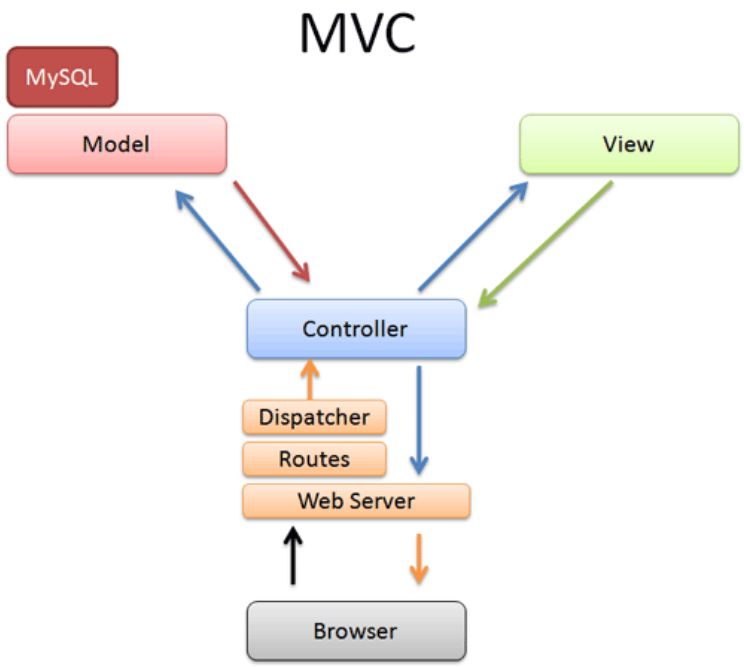


Abbildung 3.2: MVC-Modell in *CakePHP* [Bharti2009]

Das Softwaresystem wird in drei Ebenen aufgeteilt [Scherer2009]:

[2]Ruby on Rails (`http://rubyonrails.org/`) ist ein quelloffenes Web-Development Framework, das in der Programmiersprache Ruby entwickelt wird und sämtliche Design-Patterns wie MVC und Prinzipien wie "Don't repeat yourself" oder "Convention over configuration" entält.

- **Datenmodell** (englisch Model): Diese Ebene repräsentiert die Datenquelle. Datenquellen können sowohl Datenbanktabellen wie auch XML-Dateien sein. Ferner werden hier gewisse Abhängigkeiten zu anderen Daten sowie mögliche Validierungsregeln definiert.
- **Präsentation** (englisch View): Hier werden die Daten aus dem Model nach entsprechender Aufbereitung formatiert und angezeigt.
- **Controller**: Im Controller befindet sich die eigentliche Applikationslogik. Die Daten kommen aus dem Model, werden bearbeitet und an die jeweilige View weitergegeben.

Mit dieser Aufteilung ist es beispielsweise einem Designer möglich, das View-Template anzupassen ohne dass er Model oder Controller kennen muss.

Kommunikation mit Geotagger und Gazetteer

Für die Kommunikation mit dem Geotagger *geoLyzard* wurde das Python-Paket *eWrt*[3] verwendet. Das Paket kapselt den Web-Service Zugriff sowie die Verarbeitung der Tagging-Resultate und bietet somit einfaches Handling im Umgang mit *geoLyzard*. Den Geotagger selbst kann man mit verschiedenen Parametern konfigurieren. Die Gazetteer-Struktur stellt den zentralen Tuning-Parameter in dieser Arbeit dar. Tabelle 3.1 zeigt die verschiedenen Gazetteere und ihre Struktur.

[3]beziehbar von: https://svn.semanticlab.net/svn/oss/trunk/eWRT

Gazetteer	Beschreibung der Struktur
C5.000	Berücksichtigung von Städten beziehungsweise Verwaltungseinheiten $\geq$ 5.000 Einwohner
C10.000	$\geq$ 10.000 Einwohner
C100.000	$\geq$ 100.000 Einwohner
C500.000	$\geq$ 500.000 Einwohner

Tabelle 3.1: Gazetteer-Struktur als Parameter von *geoLyzard*

Wie bereits in 3.2.1 erläutert, erhält der Geotagger ein Dokument als Input. Dieses wird auf Named Entities geparst, wobei diese aufgrund eines bestimmten Fokus-Algorithmus bewertet werden. *GeoLyzard* stehen zwei Algorithmen zur Verfügung:

- "Default": Fokus-Algorithmus von Dr. Albert Weichselbraun (WU Wien) entwickelt. Dieser wird auch im Rahmen der Diplomarbeit verwendet.
- "Amitay": Fokus-Algorithmus, der auf der Funktionsweise, welche in [Amitay2004] beschrieben ist, basiert. Bei Verwendung dieses Algorithmus stehen eine Reihe weiterer Konfigurationsparameter zur Verfügung. Eine detaillierte Auflistung findet man unter: `http://www.semanticlab.net/index.php/GeoLyzard`.

Im Folgenden soll ein Beispiel die Kommunikation mit *geoLyzard* aufzeigen. Listing 3.1 zeigt den für die Kommunikation notwendigen Python-Sourcecode. Zuerst wird über die Konsole der zu verwendende Gazetteer eingelesen (Zeile 2). Dieser wird gemeinsam mit dem eingelesen File als Parameter der statischen Methode *getGeoEntities* übergeben (Zeile 7). Diese liefert wiederum ein Array von möglichen Kandidaten zurück. Um den relevantesten Eintrag aus diesen Kandidaten zu bekommen, ruft man die statische Methode *getMostRelevantEntity* auf (Zeile 12).

```
1 if __name__ == '__main__':
```

```
2    gazetteerSize = sys.argv[1]
3    TEXT = open("../tmp/testFile.txt").read()
4    print TEXT
5  #liste aller gefundenen entities
6    try:
7      res = GeoLyzard.getGeoEntities( TEXT, gazetteerSize)
8      print "\nermittelte Locations:"
9      for entity in res.values() [0]:
10       print entity
11 #ermittle das wichtigste entity
12     myCatch = GeoLyzard.getMostRelevantEntity( res )
13     print "\nErgebnis: "
14     print myCatch
15   except:
16     print "Unexpected error:", sys.exc_info()
```

Listing 3.1: Kommunikation mit *geoLyzard*

Zur Demonstration wird folgender Newsartikel aus dem BBC-Korpus an den Geotagger geschickt - dabei wird der Gazetteer C100.000 verwendet:

BBC NEWS

Marley statue unveiled in Serbia

A statue of late reggae legend Bob Marley has been unveiled in a small Serbian village during a rock festival as a token of peace in the Balkans.
Musicians from Croatia and Serbia were joined by rock fans for the midnight ceremony in Banatski Sokolac.
Organisers said Marley, who died in 1981, "promoted peace and tolerance in his music".
Serbia recently erected a statue of iconic film character Rocky, while Mostar in Bosnia has one of Bruce Lee.

Another Serbian village put up a statue to actor Johnny Weissmuller, best known for his depiction of Tarzan.
Serbian musician Jovan Matic and veteran Croatian rock star Dado Topic took the covers off the Marley statue at the Rock Village event.
It depicts him holding a guitar, with his fist raised in a defiant pose.
Organisers claim that the Marley statue, which was created by Croatian artist Davor Dukic, is the first monument in Europe to the Jamaican−born star.
Bosnia's bronze statue of Bruce Lee, erected in 2005, was seen as a symbol against ethnic divisions deepened by the country's fierce civil war in the 1990s.
The Serbian village of Zitiste put up a statue to Rocky − played by Sylvester Stallone − last year in a bid to shake off a run of bad luck.
A series of floods and landslides had led some people to believe the village was jinxed.

Listing 3.2: Beispiel-Inputdokument für *geoLyzard*

Das Ergebnis des Geo-Tagging/Fokus Prozesses sieht man hier:

ermittelte Locations:
{'entity_id': '3194828', 'name': 'Mostar', 'confidence': '8.0', 'long': '17.808055877685547', 'occurrences': '1', 'lat': '43.34333419799805', 'tree_length': '4', 'focus_points': '0.5'}
{'entity_id': '3202326', 'name': 'Republic of Croatia', 'confidence': '6.0', 'long': '15.5', 'occurrences': '1', 'lat': '45.16666793823242', 'tree_length': '2', 'focus_points': '0.0'}
{'entity_id': '3277605', 'name': 'Bosnia and Herzegovina', 'confidence': '0.0', 'long': '17.83333396911621', 'occurrences': '0', 'lat': '44.25', 'tree_length': '2', 'focus_points': '0.0'}

{'entity_id': '3229999', 'name': 'Federation of Bosnia and Herzegovina', 'confidence': '0.0', 'long': '17.58333396911621', 'occurrences': '0', 'lat': '44.0', 'tree_length': '3', 'focus_points': '0.0'}
{'entity_id': '6290252', 'name': 'Serbia', 'confidence': '6.0', 'long': '20.459976196289062', 'occurrences': '3', 'lat': '44.81892395019531', 'tree_length': '2', 'focus_points': '0.0'}
{'entity_id': '6255148', 'name': 'Europe', 'confidence': '5.0', 'long': '9.140625', 'occurrences': '1', 'lat': '48.69095993041992', 'tree_length': '1', 'focus_points': '0.5'}
{'entity_id': '2635167', 'name': 'United Kingdom of Great Britain and Northern Ireland', 'confidence': '6.0', 'long': '−4.0', 'occurrences': '1', 'lat': '54.0', 'tree_length': '2', 'focus_points': '0.0'}

Ergebnis:

{'entity_id': '3194828', 'name': 'Mostar', 'confidence': '8.0', 'long': '17.808055877685547', 'occurrences': '1', 'lat': '43.34333419799805', 'tree_length': '4', 'focus_points': '0.5'}

Listing 3.3: Output von *geoLyzard*

Wie man dem Ergebnis entnehmen kann, würde hier der Geofokus bei vorgenommener Einstellung (untere Schranke von 100.000 Einwohnern) auf die bosnische Stadt Mostar fallen. Ausschlaggebend ist hier die höhere Confidence. Bei der Algorithmus-Einstellung "Default" ist die Confidence das entscheidende Merkmal. Wenn mehrere Kandidaten die gleiche Confidence besitzen, wird aufgrund der Häufigkeit (Occurences) verglichen. Mit einem anderen Gazetteer ließen sich andere Resultate erzielen. Beispielsweise würde der Gazetteer C500.000 (ausschließlich Städte beziehungsweise Verwaltungseinheiten $\geq$ 500.000 Einwohner sind erlaubt) Mostar aufgrund der zu geringen Einwohnerzahl (ca. 110.000) ignorieren.
Weiters werden Angaben über Längen- und Breitengrad (*long*, *lat*), Anzahl der Verwaltungshierarchieebenen (*tree_length*) sowie eine eindeutige ID des Eintrags im Gazetteer (*entity_id*) gemacht.

Kommunikation mit Web-Services von Geonames

Im gezeigten Beispiel ist relativ eindeutig um "welches" Mostar es sich handelt, da dieser Name weltweit einzigartig für einen Ort ist - und zwar als Stadt in Bosnien und Herzegowina. Würde hingegen Vienna im Resultset stehen, würde ein geographisch bewanderter Mensch unschlüssig sein. Ist das berühmte Vienna in einem womöglich englisch-sprachigen Text gemeint, oder handelt es sich um einen der insgesamt 14 Namensvetter aus den USA? Auch hier bietet das Python-Paket *eWRT* Methoden um weitere Informationen über Geo-Entitäten zu ermitteln. Wie auch auf textbasierte Weise Eindeutigkeit einer Geo-Entität erreicht werden kann, zeigt Listing 3.4. Diese Methoden managen die Kommunikation mit den Web-Services von *Geonames*. Eine vollständige Liste aller verfügbaren Web-Services ist hier zu finden:
http://www.geonames.org/export/ws-overview.html.
Dabei muss lediglich ein Objekt der Klasse *Gazetteer* instanziiert werden und dessen Instanzmethode die *geoname_id* übergeben werden.

```
1 g = Gazetteer()
2
3 def getGeoUrlFromId(id):
4     return g.getGeoEntityDict( id=id )
```

Listing 3.4: Kommunikation mit *Geonames*

Die Ausgabe sieht wie folgt aus:

```
python getGeoUrlFromId.py 276136

Europe>Republic of Austria>Bundesland Wien>Wien
```

Listing 3.5: Ausgabe des Geo-Urls der *geoname_id* 276136

Die eindeutige *geoname_id* (2761369) wird dabei in den eindeutigen Geo-Url in kanonischer Form umgewandelt.
Ebenfalls beherrscht *eWrt* den umgekehrten Weg, d. h. die Umwandlung eines Geo-Urls in eine *geoname_id*.
Diese Methoden sind innerhalb der Entwicklung insbesondere beim Vergleich zwischen ermitteltem Geo-Tag und Gold-Standard Geo-Tag wichtig. Dabei wird die ermittelte *geoname_id* in einen Geo-Url umgewandelt, sodass dann dieser mit dem des Gold-Standards verglichen werden kann.

Um auf die für die ontologiebasierte Nutzenberechnung notwendigen Web-Services zuzugreifen, wurde ein *PEAR*-Paket verwendet. *PEAR*[4] ist eine Bibliothek, die Standardlösungen für Anwendungsgebiete in der Entwicklung von PHP-Applikationen anbietet. Das Paket heißt *Services_Geonames* und ist hier zu finden:
http://code.google.com/p/services-geonames/.
Verwendet wurde in diesem Zusammenhang vor allem das *neighbours* Web-Service von *Geonames*. Als Parameter erhält dieses die *geoname_id* der zu analysierenden Geo-Entität. Zurückgeliefert wird ein Array mit allen Nachbarn inklusive Namen und *geoname_ids*. Somit können semantische Zusammenhänge zu einer gewissen Geo-Entität assoziiert werden. Listing 3.6 zeigt den dafür notwendigen PHP-Code.

```
1 require_once 'Services/GeoNames.php';
2
3 $geonames = new Services_GeoNames('username', 'some authtoken...');
4 $neighbours = $geonames->neighbours(2782113);
5
6 echo ("<BR/><BR/>Neighbours of Austria:<br/>");
7 foreach ($neighbours as $neighbour)
8 {
```

[4] http://pear.php.net/

```
9     printf(" – Name: %s (geonameId: %u)<br/>", $neighbour->name,
          $neighbour->geonameId);
10 }
```

Listing 3.6: Ermittlung von Nachbarn durch das *Services_ Geonames* Paket

Es werden die Nachbarn von Österreich (*geoname_ id*=2782113) gesucht. Nachfolgend das Ergebnis dieser Anfrage:

```
Neighbours of Austria:
– Name: Czech Republic (geonameId: 3077311)
– Name: Germany (geonameId: 2921044)
– Name: Hungary (geonameId: 719819)
– Name: Italy (geonameId: 3175395)
– Name: Liechtenstein (geonameId: 3042058)
– Name: Slovakia (geonameId: 3057568)
– Name: Slovenia (geonameId: 3190538)
– Name: Switzerland (geonameId: 2658434)
```

Listing 3.7: Output der Neighbour-Anfrage

3.2.3 Programmlogik

In diesem Kapitel wird der für die Nutzenberechnung verantwortliche Algorithmus betrachtet. Als Basis hierfür dient das Paper: "A utility centered approach for evaluating and optimizing geo-tagging" von Dr. Albert Weichselbraun [Weichselbraun2009].
Der Gesamtnutzen berechnet sich aus zwei "Teilnutzen", die einerseits durch eine Hierarchie-Evaluierung (u_h) und andererseits durch eine Ontologie-Evaluierung (u_o) ermittelt werden:

$$u_{total} = u_h + u_o \tag{3.1}$$

Die hierarchiebasierte Berechnung des Nutzens ergibt sich aus der Anzahl der übereinstimmenden Hierarchieebenen in Relation zum Maximum an vorkommenden Hierarchieebenen:

$$u_h = \frac{|S_{correct} \cap S_{suggested}|}{max(|S_{correct}|, |S_{suggested}|)} \tag{3.2}$$

Idente Tags würden so u_h von eins und u_o von null generieren. Abweichungen führen zu $u_h < 1$ und $u_o \geq 0$.

Folgendes Beispiel soll diese Überlegungen verdeutlichen:

Gold-Standard Geo-Tag: **Europa/Deutschland**/Bayern/München
$\Rightarrow S_{correct} = 4$ (Hierarchieebenen)

ermitteltes Tag: **Europa/Deutschland**
$\Rightarrow S_{suggested} = 2$

$$u_h = \frac{2}{max(2,4)} \Rightarrow u_h = 0,5 \tag{3.3}$$

Die hervorgehobenen Hierarchieebenen (n_{equal}=2) zeigen die Übereinstimmung an, wobei sich die maximale Anzahl auf n_{total}=4 beläuft.

Der Teil der **ontologiebasierten** Berechnung versucht über diesen recht simplen Hierarchieebenenvergleich hinaus semantische Zusammenhänge zwischen den Tags zu erkennen. Die Berechnung definiert sich wie folgt:

$$u_o = (1 - u_h) \cdot f_{eval} \Rightarrow f_{eval} = \prod_{j=1}^{n} w_{dj} \tag{3.4}$$

Im Gegensatz zum **hierarchiebasierten** Teil hat der Benutzer hier Einfluss in Form von entsprechenden Gewichtungen auf die Berechnung von u_o. Das Ergebnis (f_{eval}) der Ontologie-Evaluierung setzt sich aus dem Produkt der benutzerspezifischen Gewichtungen (w_{d_j} [0;1]) für die jeweiligen Ontologievergleiche zusammen.

Ein Benutzer, der sich für Wandern in Ostösterreich interessiert, wird ein Dokument, das anstatt "Europa/Österreich/Niederösterreich" mit "Europa/Österreich/Steiermark" getaggt wurde wahrscheinlich nutzenbringender finden als jemand der explizit nach Winzer in der Wachau sucht.
Der Benutzer hat innerhalb des Frameworks die Möglichkeit diesen teilrichtigen Ergebnissen Nutzen zuzuweisen. Dabei muss er sich Fragen stellen wie: "Inwiefern stiftet mir ein Dokument, das anstelle einer bestimmten Stadt mit dem umliegenden Bundesland getaggt wurde noch Nutzen?" oder "Inwiefern stellt mich ein Dokument zufrieden, das statt eines bestimmten Landes mit einem Nachbarland getaggt wurde?"
In einem ersten Schritt bewertet der Benutzer die "Zufriedenheit" der Granularität eines teilrichtigen Tagging-Ergebnisses. Abbildung 3.3 zeigt eine Darstellung der verschiedenen Granularitätsausprägungen. Gemeint ist hier mit Zufriedenheit, die Akzeptanz hinsichtlich "zu genauer" (zum Beispiel anstatt des korrekten Bundeslands -> eine Stadt des Bundesland) oder "zu ungenauer" (anstatt der korrekten Stadt -> umgebendes Bundesland) Ergebnisse in einem ersten Ansatz zu parametrisieren und zu erfassen. Aus dem Aufbau des Geo-Urls mache ich mir ontologische Aspekte wie *contains*- und *partOf*-Beziehungen zu Nutze. Ein Land ist ein Teil eines Kontinents (*partOf*) und im umgekehrten Fall besteht ein Kontinent aus einem oder mehreren Ländern (*contains*).

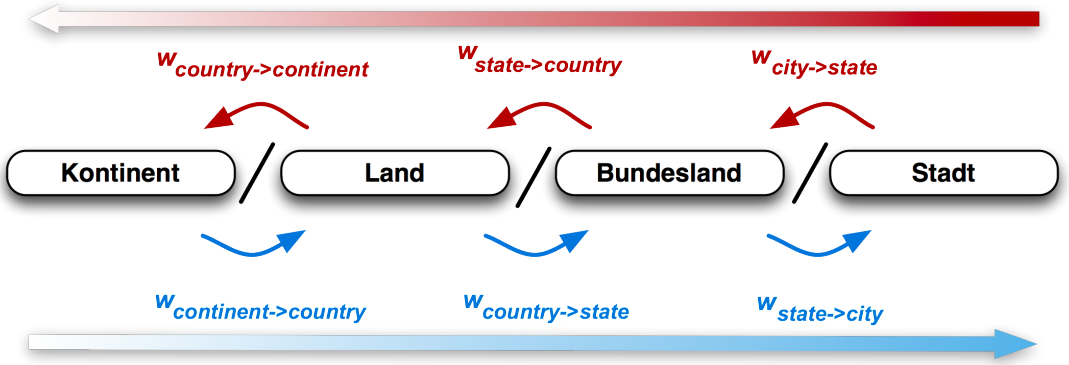


Abbildung 3.3: Bewegungen entlang der Hierarchieebenen

Aus der Grafik lässt sich erkennen, dass das Bewegen entlang der Hierarchieebenen mit Kosten verbunden ist. Die Benutzer passen diese Kosten ihren Bedürfnissen und Wünschen an. Beispielsweise würde ein teilrichtiges Ergebnis, das anstatt der Stadt das umgebende Bundesland enthält, mit dem Gewicht $w_{\text{city->state}}$ erfasst.
Der Nutzen eines teilrichtigen Ergebnisses für das gilt:

$$|S_{correct}| - |S_{suggested}| > 1; |S_{correct}| \cap |S_{suggested}| > 0$$

wird durch das Produkt der jeweiligen Gewichte, wie in Formel 3.4 dargestellt, ermittelt. Eine weitere Bedingung ist, dass alle Hierarchieebenen eines Geo-Urls komplett im Geo-Url des anderen enthalten sein müssen. Andernfalls werden durch die ontologische Evaluierung keine Punkte erzielt.
Als Beispiel dazu folgendes Benutzerprofil:

$$w_{country->continent} = 0$$
$$w_{state->country} = 0{,}7$$
$$w_{city->state} = 0{,}7$$
$$w_{continent->country} = 0$$
$$w_{country->state} = 0{,}5$$
$$w_{state->city} = 0{,}5$$

Diese Gewichtungen spiegeln lediglich die subjektive Zufriedenheit hinsichtlich der Tagging-Ergebnisse des Benutzers wieder. Es wird angenommen, dass der Benutzer, allgemeine Informationen über Süddeutschland erlangen möchte. Aufgrund dieser Ausgangslage stellen dem Benutzer "zu ungenaue" Ergebnisse zufriedener als "zu genaue". Dass ein Dokument statt mit München mit Bayern getaggt wurde, stellt ihn relativ hoch zufrieden ($w_{city->state}$ = 0,7). Hingegen wird ein Dokument, das anstatt des Tags Deutschland mit Europa gekennzeichnet wurde, bereits als "zu ungenau" empfunden und stellt somit keinen Nutzen für den Benutzer da ($w_{country->continent} = 0$).
Folgende Tags werden angenommen:

Gold-Standard Geo-Tag: Europa/Deutschland/Bayern/München
ermitteltes Tag: Europa/Deutschland

Die Berechnung für f_{eval} lautet folglich:
$f_{eval} = w_{city->state} \cdot w_{state->country}$
$\Rightarrow f_{eval} = 0,7 \cdot 0,7 \Rightarrow f_{eval} = 0,49$

Die Evaluierung mit den Gewichten ist in Abbildung 3.4 ersichtlich:

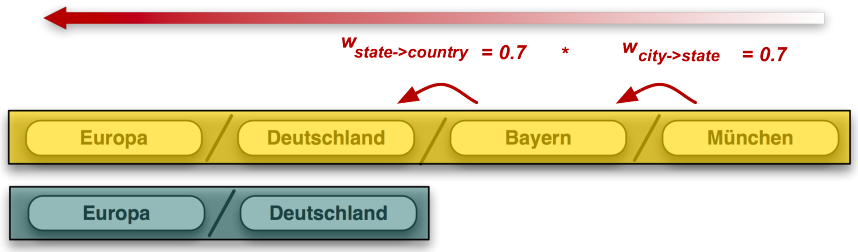


Abbildung 3.4: Evaluierung des Ontologieweges

Da das ermittelte Tag grober ist und somit weniger Hierarchieebenen als das des Gold-Standard Geo-Tags besitzt (zwei versus vier), müssen die Schritte zur letzten gemeinsamen Hierarchieebene (Deutschland) ermittelt werden.

Schlussendlich wird das Produkt dieser Schritte in Form der benutzerdefinierten Gewichte gebildet. Um nun einen ganzheitlichen Nutzen für das Tag dieses Dokuments zu berechnen, benötigt man auch den Nutzenteil aus der hierarchischen Berechnung. Dieser wurde bereits zuvor mit u_h=0.5 ermittelt. Somit gilt für die Berechnung von u_{total}:

$$u_o = (1 - u_{\text{h}}) \cdot f_{eval}$$
$$\Rightarrow u_{\text{o}} = (1 - 0,5) \cdot 0,49 \Rightarrow u_{\text{o}} = 0,245$$
$$u_{\text{total}} = u_{\text{o}} + u_{\text{h}}$$
$$\Rightarrow u_{\text{total}} = 0,745$$

Passt ein Geo-Url nicht zur Gänze in den zu vergleichenden Geo-Url, wird versucht über weitere Ontologien semantische Zusammenhänge zu finden. Im entwickelten Test-Framework konkretisiert sich dies durch Vergleiche auf Nachbars- und Nachbars-Nachbarn-Ebene. Das zu evaluierende Tag wird mit dem Gold-Standard Geo-Tag auf eine mögliche Nachbarsbeziehung geprüft. Bei zwei zu vergleichenden Städten wird zusätzlich darauf geachtet, ob diese im selben Bundesland liegen. Gegeben sei folgendes Benutzerprofil:

$$w_{country->continent} = 0$$
$$w_{state->country} = 0,7$$
$$w_{city->state} = 0,7$$
$$w_{continent->country} = 0$$
$$w_{country->state} = 0,5$$
$$w_{state->city} = 0,5$$
$$w_{isNeighbourOf} = 0,6$$
$$w_{isNeighbourOfNeighbour} = 0,3$$

und folgende Tags:

Gold-Standard Tag: Europa/Deutschland/Bayern/München
ermitteltes Tag: Europa/Deutschland/Hessen/Frankfurt

Die Berechnung kann in Abbildung 3.5 nachvollzogen werden. Zunächst werden die Kosten von München nach Bayern ermittelt ($w_{city->state} = 0,7$),

danach kann die Nachbarsbeziehung hergestellt werden ($w_{isNeighbourOf} = 0{,}6$). Zuletzt wird dieses Produkt mit den Kosten für die Bewegung von Hessen zu Frankfurt multipliziert ($w_{state->city} = 0{,}5$).

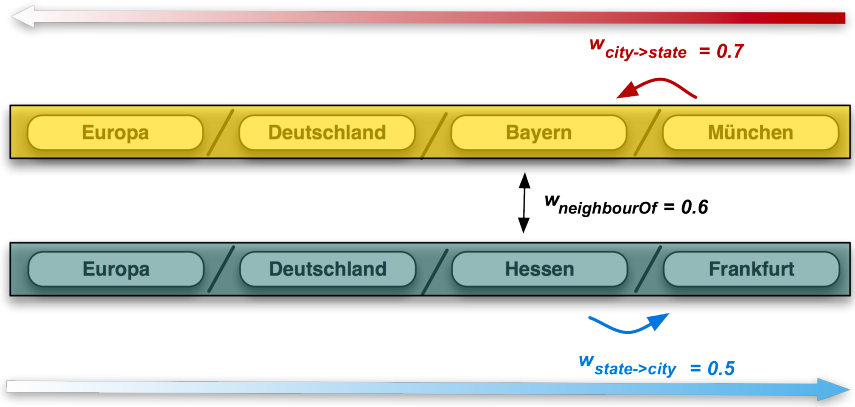


Abbildung 3.5: Evaluierung des Ontologieweges mit Nachbarsbeziehung

f_{eval} ist demnach:

$$f_{eval} = w_{city->state} \cdot w_{neighbourOf} \cdot w_{state->city}$$
$$\Rightarrow f_{eval} = 0,7 \cdot 0,6 \cdot 0,5 \Rightarrow f_{eval} = 0,21$$

Eine Übersicht mit allen verfügbaren benutzerspezifischen Gewichten ist in Tabelle 3.2 dargestellt.

Gewicht	Beschreibung
$w_{city\text{->}state}$	**statt** korrektem **Ort** ist umgebendes **Bundesland** Fokus
$w_{state\text{->}country}$	**statt** korrektem **Bundesland** ist umgebender **Staat** Fokus
$w_{country\text{->}continent}$	**statt** korrektem **Staat** ist sein **Kontinent** Fokus
$w_{continent\text{->}country}$	**statt** korrektem **Kontinent** ist zum Kontinent gehöriger **Staat** Fokus
$w_{country\text{->}state}$	**statt** des korrektem **Staat** ist zum Staat gehöriges **Bundesland** Fokus
$w_{state\text{->}city}$	**statt** korrektem **Bundesland** ist zum Bundesland gehöriger **Ort** Fokus
$w_{sameState}$	**statt** korrektem **Ort** ist **zum selben Bundesland gehöriger Ort** Fokus
$w_{neighbourOf}$	**statt** korrektem **Bundesland/Staat** ist **Nachbar-Bundesland/Staat** Fokus
$w_{neighbourOfNeighbour}$	statt korrektem **Bundesland/Staat** ist **Nachbar-Bundesland/Staat eines *neighbourOf*** Fokus

Tabelle 3.2: Liste der anpassbaren benutzerspezifischen Gewichte im Test-Framework

3.3 Usecases

Für die Evaluierung durch das Test-Framework wurden zwei Usecases erzeugt.

- **Usecase "Wander-Reiseführer für Niederösterreich"**: Konzentration auf ein Bundesland -> fein-granulare Ergebnisse von Bedeutung
- **Usecase "Europa-Reiseführer"**: Konzentration auf Kontinent, sowie enthaltene Länder -> grob-granulare Ergebnisse von Bedeutung

In diesen Usecases wird durch die definierten benutzerspezifischen Gewichten beschrieben, wie teilrichtige Tagging-Ergebnisse bewertet werden sollen.

Um das volle Potential der benutzerspezifischen Gewichte aufzuzeigen, werden noch zwei (wenn auch nicht realistische) Usecases hinzugefügt.

- **Usecase "no Ontology"**: benutzerspezifische Gewichte werden ignoriert ($w_* = 0$)
- **Usecase "max Ontology"**: benutzerspezifische Gewichte haben maximale Ausprägung ($w_* = 1$)

3.3.1 Erster Usecase - "Wander-Reiseführer für ein Bundesland"

Wichtig hierbei ist die Konzentration auf ein Bundesland (Niederösterreich). Zu diesem Zweck wurde ein Benutzerprofil mit folgenden benutzerspezifischen Gewichtungen erstellt:

Gewicht	Gewichtung
$w_{country->continent}$	0
$w_{state->country}$	0
$w_{city->state}$	0,8
$w_{continent->country}$	0
$w_{country->state}$	0
$w_{state->city}$	0,6
$w_{sameState}$	0,8
$w_{isNeighbourOf}$	0
$w_{isNeighbourOfNeighbour}$	0

Tabelle 3.3: Benutzerprofil für Usecase "Wander-Reiseführer"

Wie man Tabelle 3.3 entnehmen kann, ist der Benutzer auf fein-granulare Ergebnisse fokussiert. Wird anstatt eines richtigen Orts das umliegende Bundeslands getaggt, ist der Benutzer relativ hoch ($w_{city->state} = 0{,}8$) zufrieden. Das gleiche gilt auch für einen anderen Ort im selben Bundesland ($w_{sameState} = 0{,}8$). Ebenfalls liefert dem Benutzer ein Ort, der im eigentlich richtig getaggten Bundesland liegt Nutzen ($w_{state->city} = 0{,}6$) Der Benutzer ist daher möglichst auf detaillierte Informationen angewiesen.
Wird ein Dokument statt des Bundeslands mit dessen Land (zum Beispiel Österreich) getaggt, ist dies für den Benutzer uninteressant, da er damit "zu allgemeine" Informationen erhält. Wandern in den Nachbarbundesländer (zum Beispiel Steiermark) und in den Nachbarsnachbarn-Bundesländer (zum Beispiel Kärnten via Steiermark) stellt für ihn ebenfalls keinen Nutzen dar, da ihn nur Informationen über ein konkretes Bundesland (Niederösterreich) interessieren.

3.3.2 Zweiter Usecase - "Europa-Reiseführer"

Das Augenmerk bei teilrichtigen Ergebnissen ist hier auf grob-granulare Ergebnisse gerichtet. Der Benutzer möchte allgemeine und eher oberflächliche Informationen auf Country/Continent-Ebene erhalten. Er ist relativ

hoch zufrieden ($w_{country->continent} = 0{,}8$) wenn anstatt des korrekten Taggings durch ein Land (zum Beispiel Österreich) das Dokument mit dessen Kontinent (zum Beispiel Europa) getaggt wird. Der Benutzer ist auch an Informationen über Nachbarländer ($w_{isNeighbourOf} = 0{,}7$) beziehungsweise Nachbarsnachbar-Länder interessiert ($w_{isNeighbourOfNeighbour} = 0{,}5$).
Zu fein-granulare Ergebnisse (zum Beispiel St. Pölten statt Niederösterreich) haben für ihn sehr geringe Bedeutung.
Das vollständige Profil ist in Tabelle 3.4 ersichtlich:

Gewicht	**Gewichtung**
$w_{country->continent}$	0,8
$w_{state->country}$	0,4
$w_{city->state}$	0,3
$w_{continent->country}$	0,7
$w_{country->state}$	0,1
$w_{state->city}$	0,1
$w_{sameState}$	0,4
$w_{isNeighbourOf}$	0,7
$w_{isNeighbourOfNeighbour}$	0,5

Tabelle 3.4: Benutzerprofil für Usecase ”Europa-Reiseführer”

3.3.3 Weitere Usecases

Um das gesamte Einfluss-Potential der benutzerspezifischen Gewichte zu dokumentieren, werden zwei weitere Usecases definiert. Diese Usecases werden gesondert von den anderen beiden analysiert und dienen demnach nur dem Zweck das Potential der vorgestellten Methode aufzuzeigen.
Usecase ”no Ontology” ignoriert die Evaluierung anhand von Ontologien komplett (das heißt $u_o = 0$) und lässt somit lediglich den hierarchiebasierten Vergleich zu (u_h). Tabelle 3.5 zeigt das vollständige Profil von Usecase ”no Ontology”:

Gewicht	Gewichtung
$w_{country->continent}$	0
$w_{state->country}$	0
$w_{city->state}$	0
$w_{continent->country}$	0
$w_{country->state}$	0
$w_{state->city}$	0
$w_{sameState}$	0
$w_{isNeighbourOf}$	0
$w_{isNeighbourOfNeighbour}$	0

Tabelle 3.5: Benutzerprofil für Usecase ”no Ontology”

Usecase ”max Ontology” adressiert genau das Gegenteil indem allen benutzerspezifischen Gewichten das Maximum zugewiesen wird. Das vollständige Profil zeigt Tabelle 3.6.

Gewicht	Gewichtung
$w_{country->continent}$	1
$w_{state->country}$	1
$w_{city->state}$	1
$w_{continent->country}$	1
$w_{country->state}$	1
$w_{state->city}$	1
$w_{sameState}$	1
$w_{isNeighbourOf}$	1
$w_{isNeighbourOfNeighbour}$	1

Tabelle 3.6: Benutzerprofil für Usecase ”max Ontology”

3.4 Ergebnisse

Die in 3.3 vorgestellten Usecases werden mit folgenden Konfigurationen getestet:

- Geotagger: *GeoLyzard* (Algorithmus: "Default")
- Gazetteer: *Geonames*
- verwendete Gazetteers:
  - C5.000 (= Geo-Entitäten ab einer Einwohnerzahl von 5.000)
  - C100.000
  - C500.000
- Korpus: 500 BBC-Newsartikel

Eine detailliertere Auflistung (Anzahl der Entitäten) der einzelnen Gazetteers zeigt Tabelle 3.7.

Gazetteer-Name	**min. Einwohnerzahl**	**Anzahl Entitäten**
C5.000	5.000	115.184
C100.000	100.000	28.492
C500.000	500.000	17.051

Tabelle 3.7: Überblick über die verwendeten Gazetteers

Ergebnisse mit Gazetteer C5.000

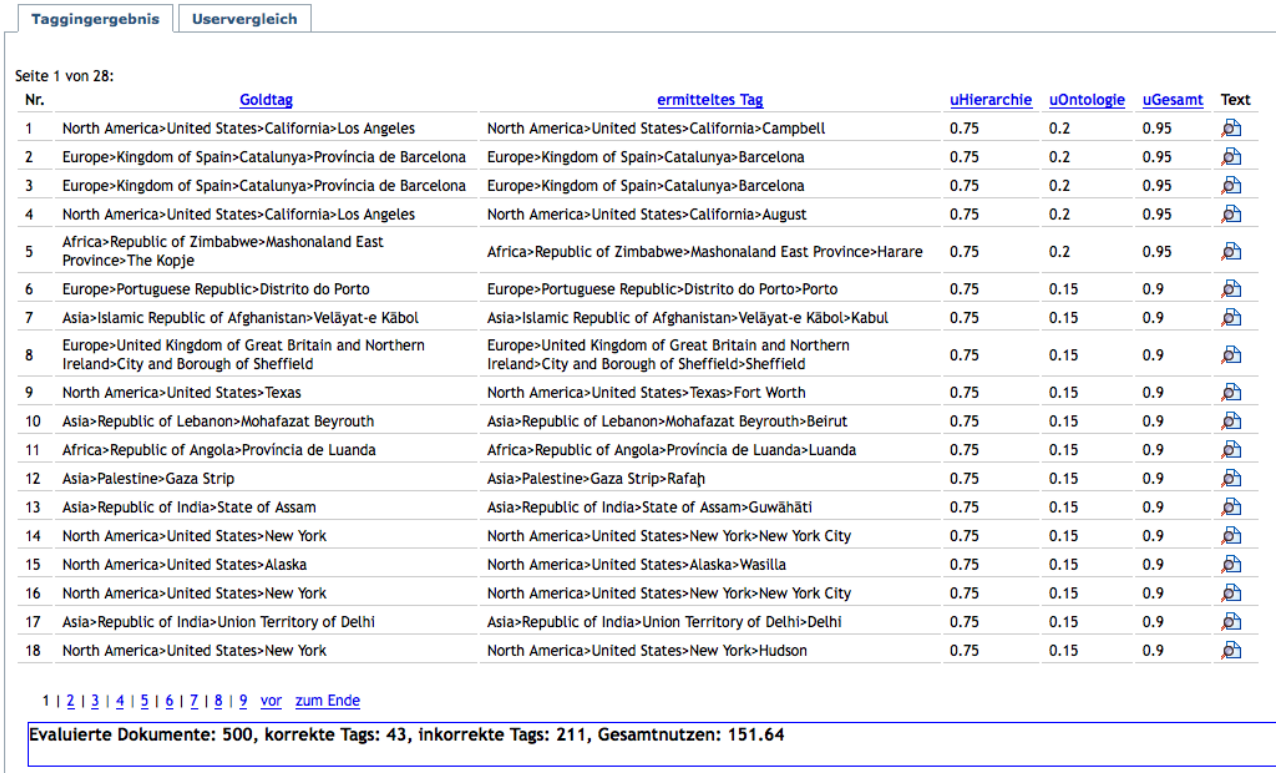
Tabelle 3.8 zeigt das Ergebnis der Evaluierung von 500 Newsartikeln durch den Gazetteer C5.000, der Geo-Entitäten $\geq$ 5.000 Einwohner enthält. Während 43 Dokumente richtig (u_{total}=1) getaggt wurden, wurde bei 211 Dokumenten kein Nutzen (u_{total}=0) erzielt. Festzustellen ist, dass trotz stark unterschiedlicher Konfiguration beide Benutzerprofile sehr ähnlichen Nutzen liefern (u_{total}=151,64 vs. u_{total}=149,09). Diese geringe Differenz kommt aufgrund der geringen Anzahl an Ontologie-Evaluierungen und der niedrigen Gewichtung für fein-granulare Resultate des zweiten Usecases ("Europa-Reiseführer") zustande.
Welch großer Nutzen-Unterschied möglich ist, zeigt der Vergleich der zwei konträren Usecases "no Ontology" und "max Ontology". Usecase "max Ontology" generiert mehr als den 1,5 fachen Nutzen als Usecase "no Ontology",

da dieser lediglich Nutzen aus dem hierarchiebasierten Vergleich zieht (AnzOnt=0).

Usecase	korrekt	inkorrekt	Gesamtnutzen	Anzahl Ont-Eval
Wander-Reiseführer	43	211	151.64	22
Europa-Reiseführer	43	211	150,53	162
no Ontology	43	211	148,09	0
max Ontology	43	211	230,49	162

Tabelle 3.8: Geotagging-Ergebnis mit Gazetteer C5.000

Abbildungen 3.6 und 3.7 zeigen die Evaluierungsergebnisse nach Ontologie-Punkten sortiert. Da Usecase zwei, das breiter gestreute Profil besitzt, erzielt dieser weniger Ontologie-Punkte. Grund dafür ist die geringe Gewichtung für fein-granulare Ergebnisse. Gerade bei diesem Gazetteer (Geo-Entitäten bereits ab 5.000 Einwohner enthalten) ist zu erwarten, dass tendenziell feinere Ergebnisse erzielt werden. Usecase eins erzielt zwar nur bei 22 Dokumenten Ontologiepunkte, dafür ist aber die starke Gewichtung von $w_{sameState}$ und $w_{state->city}$ für den im Vergleich zu Usecase zwei höheren Gesamtnutzen verantwortlich.



Taggingergebnis | Uservergleich

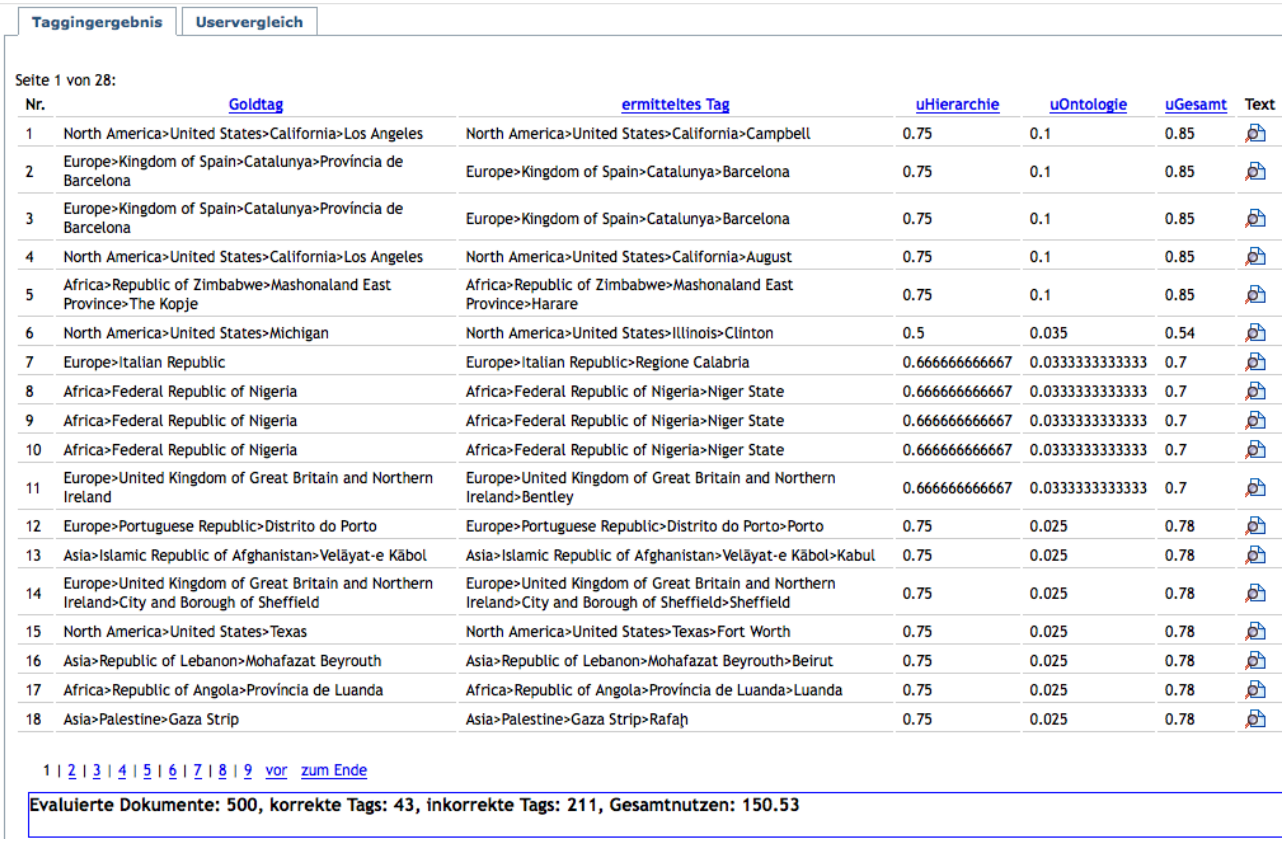
Seite 1 von 28:

Nr.	Goldtag	ermitteltes Tag	uHierarchie	uOntologie	uGesamt	Text
1	North America>United States>California>Los Angeles	North America>United States>California>Campbell	0.75	0.2	0.95	
2	Europe>Kingdom of Spain>Catalunya>Província de Barcelona	Europe>Kingdom of Spain>Catalunya>Barcelona	0.75	0.2	0.95	
3	Europe>Kingdom of Spain>Catalunya>Província de Barcelona	Europe>Kingdom of Spain>Catalunya>Barcelona	0.75	0.2	0.95	
4	North America>United States>California>Los Angeles	North America>United States>California>August	0.75	0.2	0.95	
5	Africa>Republic of Zimbabwe>Mashonaland East Province>The Kopje	Africa>Republic of Zimbabwe>Mashonaland East Province>Harare	0.75	0.2	0.95	
6	Europe>Portuguese Republic>Distrito do Porto	Europe>Portuguese Republic>Distrito do Porto>Porto	0.75	0.15	0.9	
7	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol>Kabul	0.75	0.15	0.9	
8	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Sheffield	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Sheffield>Sheffield	0.75	0.15	0.9	
9	North America>United States>Texas	North America>United States>Texas>Fort Worth	0.75	0.15	0.9	
10	Asia>Republic of Lebanon>Mohafazat Beyrouth	Asia>Republic of Lebanon>Mohafazat Beyrouth>Beirut	0.75	0.15	0.9	
11	Africa>Republic of Angola>Província de Luanda	Africa>Republic of Angola>Província de Luanda>Luanda	0.75	0.15	0.9	
12	Asia>Palestine>Gaza Strip	Asia>Palestine>Gaza Strip>Rafaḥ	0.75	0.15	0.9	
13	Asia>Republic of India>State of Assam	Asia>Republic of India>State of Assam>Guwāhāti	0.75	0.15	0.9	
14	North America>United States>New York	North America>United States>New York>New York City	0.75	0.15	0.9	
15	North America>United States>Alaska	North America>United States>Alaska>Wasilla	0.75	0.15	0.9	
16	North America>United States>New York	North America>United States>New York>New York City	0.75	0.15	0.9	
17	Asia>Republic of India>Union Territory of Delhi	Asia>Republic of India>Union Territory of Delhi>Delhi	0.75	0.15	0.9	
18	North America>United States>New York	North America>United States>New York>Hudson	0.75	0.15	0.9	

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 vor zum Ende

Evaluierte Dokumente: 500, korrekte Tags: 43, inkorrekte Tags: 211, Gesamtnutzen: 151.64

Abbildung 3.6: Evaluierungsergebnisse nach Ontologie-Punkten sortiert ("Wander-Reiseführer"), Gazetteer C5.000



Taggingergebnis | Uservergleich

Seite 1 von 28:

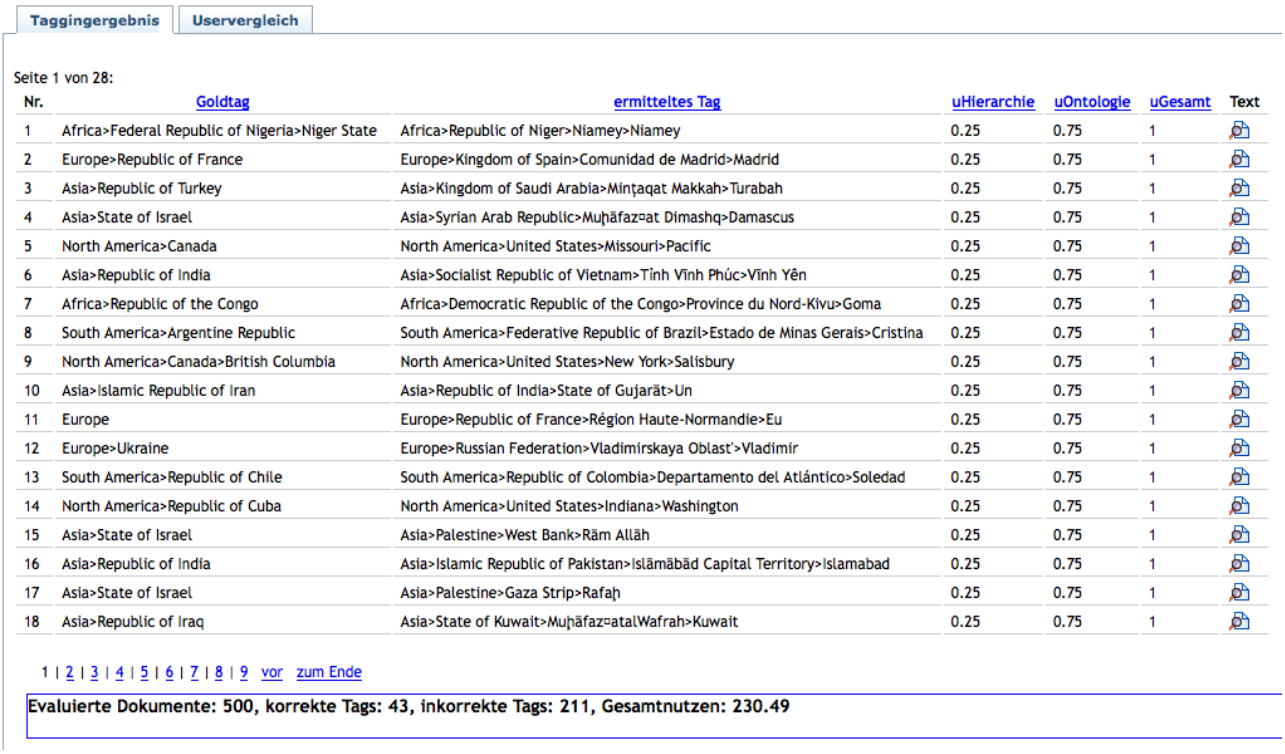
Nr.	Goldtag	ermitteltes Tag	uHierarchie	uOntologie	uGesamt	Text
1	North America>United States>California>Los Angeles	North America>United States>California>Campbell	0.75	0.1	0.85	
2	Europe>Kingdom of Spain>Catalunya>Província de Barcelona	Europe>Kingdom of Spain>Catalunya>Barcelona	0.75	0.1	0.85	
3	Europe>Kingdom of Spain>Catalunya>Província de Barcelona	Europe>Kingdom of Spain>Catalunya>Barcelona	0.75	0.1	0.85	
4	North America>United States>California>Los Angeles	North America>United States>California>August	0.75	0.1	0.85	
5	Africa>Republic of Zimbabwe>Mashonaland East Province>The Kopje	Africa>Republic of Zimbabwe>Mashonaland East Province>Harare	0.75	0.1	0.85	
6	North America>United States>Michigan	North America>United States>Illinois>Clinton	0.5	0.035	0.54	
7	Europe>Italian Republic	Europe>Italian Republic>Regione Calabria	0.666666666667	0.0333333333333	0.7	
8	Africa>Federal Republic of Nigeria	Africa>Federal Republic of Nigeria>Niger State	0.666666666667	0.0333333333333	0.7	
9	Africa>Federal Republic of Nigeria	Africa>Federal Republic of Nigeria>Niger State	0.666666666667	0.0333333333333	0.7	
10	Africa>Federal Republic of Nigeria	Africa>Federal Republic of Nigeria>Niger State	0.666666666667	0.0333333333333	0.7	
11	Europe>United Kingdom of Great Britain and Northern Ireland	Europe>United Kingdom of Great Britain and Northern Ireland>Bentley	0.666666666667	0.0333333333333	0.7	
12	Europe>Portuguese Republic>Distrito do Porto	Europe>Portuguese Republic>Distrito do Porto>Porto	0.75	0.025	0.78	
13	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol>Kabul	0.75	0.025	0.78	
14	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Sheffield	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Sheffield>Sheffield	0.75	0.025	0.78	
15	North America>United States>Texas	North America>United States>Texas>Fort Worth	0.75	0.025	0.78	
16	Asia>Republic of Lebanon>Mohafazat Beyrouth	Asia>Republic of Lebanon>Mohafazat Beyrouth>Beirut	0.75	0.025	0.78	
17	Africa>Republic of Angola>Província de Luanda	Africa>Republic of Angola>Província de Luanda>Luanda	0.75	0.025	0.78	
18	Asia>Palestine>Gaza Strip	Asia>Palestine>Gaza Strip>Rafaḥ	0.75	0.025	0.78	

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 vor zum Ende

Evaluierte Dokumente: 500, korrekte Tags: 43, inkorrekte Tags: 211, Gesamtnutzen: 150.53

Abbildung 3.7: Evaluierungsergebnisse nach Ontologie-Punkten sortiert (”Europa-Reiseführer”), Gazetteer C5.000

Um speziell die Ergebnisse der Ontologie-Evaluierung hervorzuheben, ist in Abbildung 3.8 das Evaluierungsergebnis von Usecase ”max Ontology” dargestellt. Zeile 2 zeigt ein Ontologie-Ergebnis von 0,75, das sich aus den Gewichten (alle 1) $w_{neighbourOf}$ * $w_{country->state}$ * $w_{state->city}$ multipliziert mit dem Restnutzen (=0,75) ergibt. Spanien ist ein Nachbar von Frankreich ($\mathrm{w_{neighbourOf}}$), die Region Comunidad de Madrid ist ein Teil Spaniens ($w_{country->state}$) und Madrid liegt wiederum in der autonomen Region Comunidad de Madrid ($w_{state->city}$).
Auf die Darstellung der Evaluierungsergebnisse des Usecase ”no Ontology” wird verzichtet, da keine Ontologie-Evaluierung stattfindet.



Taggingergebnis | Uservergleich

Seite 1 von 28:

Nr.	Goldtag	ermitteltes Tag	uHierarchie	uOntologie	uGesamt	Text
1	Africa>Federal Republic of Nigeria>Niger State	Africa>Republic of Niger>Niamey>Niamey	0.25	0.75	1	
2	Europe>Republic of France	Europe>Kingdom of Spain>Comunidad de Madrid>Madrid	0.25	0.75	1	
3	Asia>Republic of Turkey	Asia>Kingdom of Saudi Arabia>Minţaqat Makkah>Turabah	0.25	0.75	1	
4	Asia>State of Israel	Asia>Syrian Arab Republic>Muḥāfaz¤at Dimashq>Damascus	0.25	0.75	1	
5	North America>Canada	North America>United States>Missouri>Pacific	0.25	0.75	1	
6	Asia>Republic of India	Asia>Socialist Republic of Vietnam>Tỉnh Vĩnh Phúc>Vĩnh Yên	0.25	0.75	1	
7	Africa>Republic of the Congo	Africa>Democratic Republic of the Congo>Province du Nord-Kivu>Goma	0.25	0.75	1	
8	South America>Argentine Republic	South America>Federative Republic of Brazil>Estado de Minas Gerais>Cristina	0.25	0.75	1	
9	North America>Canada>British Columbia	North America>United States>New York>Salisbury	0.25	0.75	1	
10	Asia>Islamic Republic of Iran	Asia>Republic of India>State of Gujarāt>Un	0.25	0.75	1	
11	Europe	Europe>Republic of France>Région Haute-Normandie>Eu	0.25	0.75	1	
12	Europe>Ukraine	Europe>Russian Federation>Vladimirskaya Oblast'>Vladimir	0.25	0.75	1	
13	South America>Republic of Chile	South America>Republic of Colombia>Departamento del Atlántico>Soledad	0.25	0.75	1	
14	North America>Republic of Cuba	North America>United States>Indiana>Washington	0.25	0.75	1	
15	Asia>State of Israel	Asia>Palestine>West Bank>Rām Allāh	0.25	0.75	1	
16	Asia>Republic of India	Asia>Islamic Republic of Pakistan>Islāmābād Capital Territory>Islamabad	0.25	0.75	1	
17	Asia>State of Israel	Asia>Palestine>Gaza Strip>Rafaḥ	0.25	0.75	1	
18	Asia>Republic of Iraq	Asia>State of Kuwait>Muḥāfaz¤atalWafrah>Kuwait	0.25	0.75	1	

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 vor zum Ende

Evaluierte Dokumente: 500, korrekte Tags: 43, inkorrekte Tags: 211, Gesamtnutzen: 230.49

Abbildung 3.8: Evaluierungsergebnisse nach Ontologie-Punkten sortiert (”max Ontology”), Gazetteer C5.000

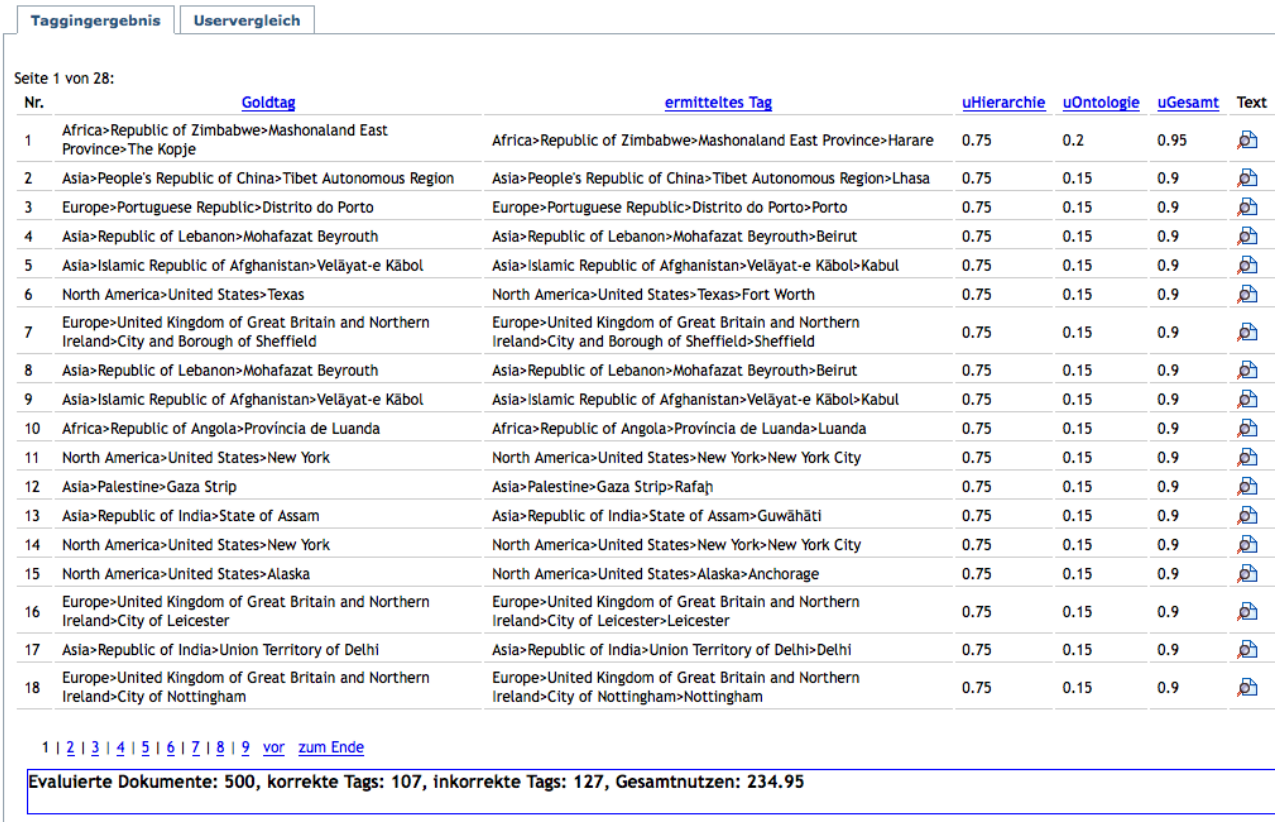
Ergebnisse mit Gazetteer C100.000

Tabelle 3.9 zeigt das Ergebnis der Evaluierung von 500 Newsartikeln durch einen Gazetteer, der Geo-Entitäten $\geq$ 100.000 Einwohner enthält. Durch die Erhöhung dieser Schranke (von Geo-Entitäten $\geq$ 5.000 auf Geo-Entitäten $\geq$ 100.000) kommt es zu großen Veränderungen. Aufgrund der Tatsache, dass die Gold-Standard Geo-Tags relativ grob sind (zumeist auf Country-Ebene), wirkt sich die Erhöhung der Minimum-Schranke positiv auf die Anzahl richtig getaggter Newsartikel aus. Waren bei Gazetteer C5.000 noch 8,6% richtig getaggt, sind hier bereits mehr als ein Fünftel (21,4%) richtig. Ähnlich stark verbessert sich der Gesamtnutzen für die Usecases ”no Ontology” (Verbesserung um 156%) und ”max Ontology” (Verbesserung um 144%) nach dem Umstieg auf Gazetteer C100.000.

Usecase	korrekt	inkorrekt	Gesamtnutzen	Anzahl Ont-Eval
Wander-Reiseführer	107	127	234.95	22
Europa-Reiseführer	107	127	235,34	203
no Ontology	107	127	231,45	0
max Ontology	107	127	331,72	203

Tabelle 3.9: Geotagging-Ergebnis mit Gazetteer C100.000

Abbildungen 3.9 und 3.10 zeigen die Evaluierungsergebnisse beider Usecases nach Ontologie-Punkten sortiert. Es zeigt sich, dass die nunmehr grober ermittelten Tags Usecase zwei ("Europa-Reiseführer"), im Gegensatz zu Usecase eins ("Wander-Reiseführer") stärker positiv beeinflussen. Der Gesamtnutzen von Usecase zwei konnte jenen von Usecase eins überholen (u_{total}=235,34 vs. u_{total}=234,95). Konnten bei Gazetteer C5.000 noch 7,4 mal (162 vs. 22) soviele Ontologie-Ermittlungen bei Usecase zwei registriert werden, erhöht sich dieses Verhältnis bei Verwendung des Gazetteers C100.000 auf 8,43 (203 vs. 22).



Taggingergebnis | Uservergleich

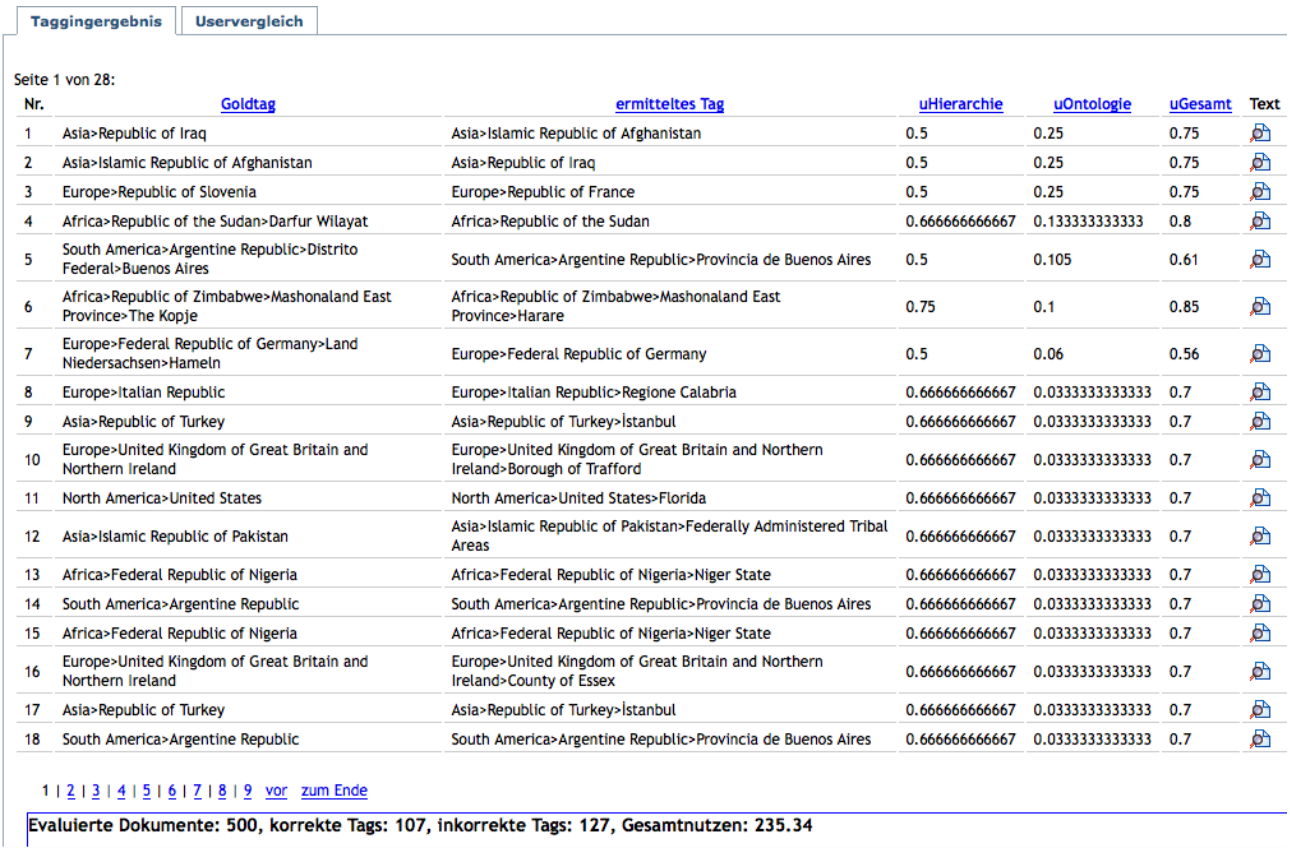
Seite 1 von 28:

Nr.	Goldtag	ermitteltes Tag	uHierarchie	uOntologie	uGesamt	Text
1	Africa>Republic of Zimbabwe>Mashonaland East Province>The Kopje	Africa>Republic of Zimbabwe>Mashonaland East Province>Harare	0.75	0.2	0.95	
2	Asia>People's Republic of China>Tibet Autonomous Region	Asia>People's Republic of China>Tibet Autonomous Region>Lhasa	0.75	0.15	0.9	
3	Europe>Portuguese Republic>Distrito do Porto	Europe>Portuguese Republic>Distrito do Porto>Porto	0.75	0.15	0.9	
4	Asia>Republic of Lebanon>Mohafazat Beyrouth	Asia>Republic of Lebanon>Mohafazat Beyrouth>Beirut	0.75	0.15	0.9	
5	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol>Kabul	0.75	0.15	0.9	
6	North America>United States>Texas	North America>United States>Texas>Fort Worth	0.75	0.15	0.9	
7	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Sheffield	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Sheffield>Sheffield	0.75	0.15	0.9	
8	Asia>Republic of Lebanon>Mohafazat Beyrouth	Asia>Republic of Lebanon>Mohafazat Beyrouth>Beirut	0.75	0.15	0.9	
9	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol>Kabul	0.75	0.15	0.9	
10	Africa>Republic of Angola>Província de Luanda	Africa>Republic of Angola>Província de Luanda>Luanda	0.75	0.15	0.9	
11	North America>United States>New York	North America>United States>New York>New York City	0.75	0.15	0.9	
12	Asia>Palestine>Gaza Strip	Asia>Palestine>Gaza Strip>Rafaḥ	0.75	0.15	0.9	
13	Asia>Republic of India>State of Assam	Asia>Republic of India>State of Assam>Guwāhāti	0.75	0.15	0.9	
14	North America>United States>New York	North America>United States>New York>New York City	0.75	0.15	0.9	
15	North America>United States>Alaska	North America>United States>Alaska>Anchorage	0.75	0.15	0.9	
16	Europe>United Kingdom of Great Britain and Northern Ireland>City of Leicester	Europe>United Kingdom of Great Britain and Northern Ireland>City of Leicester>Leicester	0.75	0.15	0.9	
17	Asia>Republic of India>Union Territory of Delhi	Asia>Republic of India>Union Territory of Delhi>Delhi	0.75	0.15	0.9	
18	Europe>United Kingdom of Great Britain and Northern Ireland>City of Nottingham	Europe>United Kingdom of Great Britain and Northern Ireland>City of Nottingham>Nottingham	0.75	0.15	0.9	

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 vor zum Ende

Evaluierte Dokumente: 500, korrekte Tags: 107, inkorrekte Tags: 127, Gesamtnutzen: 234.95

Abbildung 3.9: Evaluierungsergebnisse nach Ontologie-Punkten sortiert ("Wander-Reiseführer"), Gazetteer C100.000



Taggingergebnis | Uservergleich

Seite 1 von 28:

Nr.	Goldtag	ermitteltes Tag	uHierarchie	uOntologie	uGesamt	Text
1	Asia>Republic of Iraq	Asia>Islamic Republic of Afghanistan	0.5	0.25	0.75	
2	Asia>Islamic Republic of Afghanistan	Asia>Republic of Iraq	0.5	0.25	0.75	
3	Europe>Republic of Slovenia	Europe>Republic of France	0.5	0.25	0.75	
4	Africa>Republic of the Sudan>Darfur Wilayat	Africa>Republic of the Sudan	0.666666666667	0.133333333333	0.8	
5	South America>Argentine Republic>Distrito Federal>Buenos Aires	South America>Argentine Republic>Provincia de Buenos Aires	0.5	0.105	0.61	
6	Africa>Republic of Zimbabwe>Mashonaland East Province>The Kopje	Africa>Republic of Zimbabwe>Mashonaland East Province>Harare	0.75	0.1	0.85	
7	Europe>Federal Republic of Germany>Land Niedersachsen>Hameln	Europe>Federal Republic of Germany	0.5	0.06	0.56	
8	Europe>Italian Republic	Europe>Italian Republic>Regione Calabria	0.666666666667	0.0333333333333	0.7	
9	Asia>Republic of Turkey	Asia>Republic of Turkey>İstanbul	0.666666666667	0.0333333333333	0.7	
10	Europe>United Kingdom of Great Britain and Northern Ireland	Europe>United Kingdom of Great Britain and Northern Ireland>Borough of Trafford	0.666666666667	0.0333333333333	0.7	
11	North America>United States	North America>United States>Florida	0.666666666667	0.0333333333333	0.7	
12	Asia>Islamic Republic of Pakistan	Asia>Islamic Republic of Pakistan>Federally Administered Tribal Areas	0.666666666667	0.0333333333333	0.7	
13	Africa>Federal Republic of Nigeria	Africa>Federal Republic of Nigeria>Niger State	0.666666666667	0.0333333333333	0.7	
14	South America>Argentine Republic	South America>Argentine Republic>Provincia de Buenos Aires	0.666666666667	0.0333333333333	0.7	
15	Africa>Federal Republic of Nigeria	Africa>Federal Republic of Nigeria>Niger State	0.666666666667	0.0333333333333	0.7	
16	Europe>United Kingdom of Great Britain and Northern Ireland	Europe>United Kingdom of Great Britain and Northern Ireland>County of Essex	0.666666666667	0.0333333333333	0.7	
17	Asia>Republic of Turkey	Asia>Republic of Turkey>İstanbul	0.666666666667	0.0333333333333	0.7	
18	South America>Argentine Republic	South America>Argentine Republic>Provincia de Buenos Aires	0.666666666667	0.0333333333333	0.7	

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 vor zum Ende

Evaluierte Dokumente: 500, korrekte Tags: 107, inkorrekte Tags: 127, Gesamtnutzen: 235.34

Abbildung 3.10: Evaluierungsergebnisse nach Ontologie-Punkten sortiert ("Europa-Reiseführer"), Gazetteer C100.000

Ergebnisse mit Gazetteer C500.000

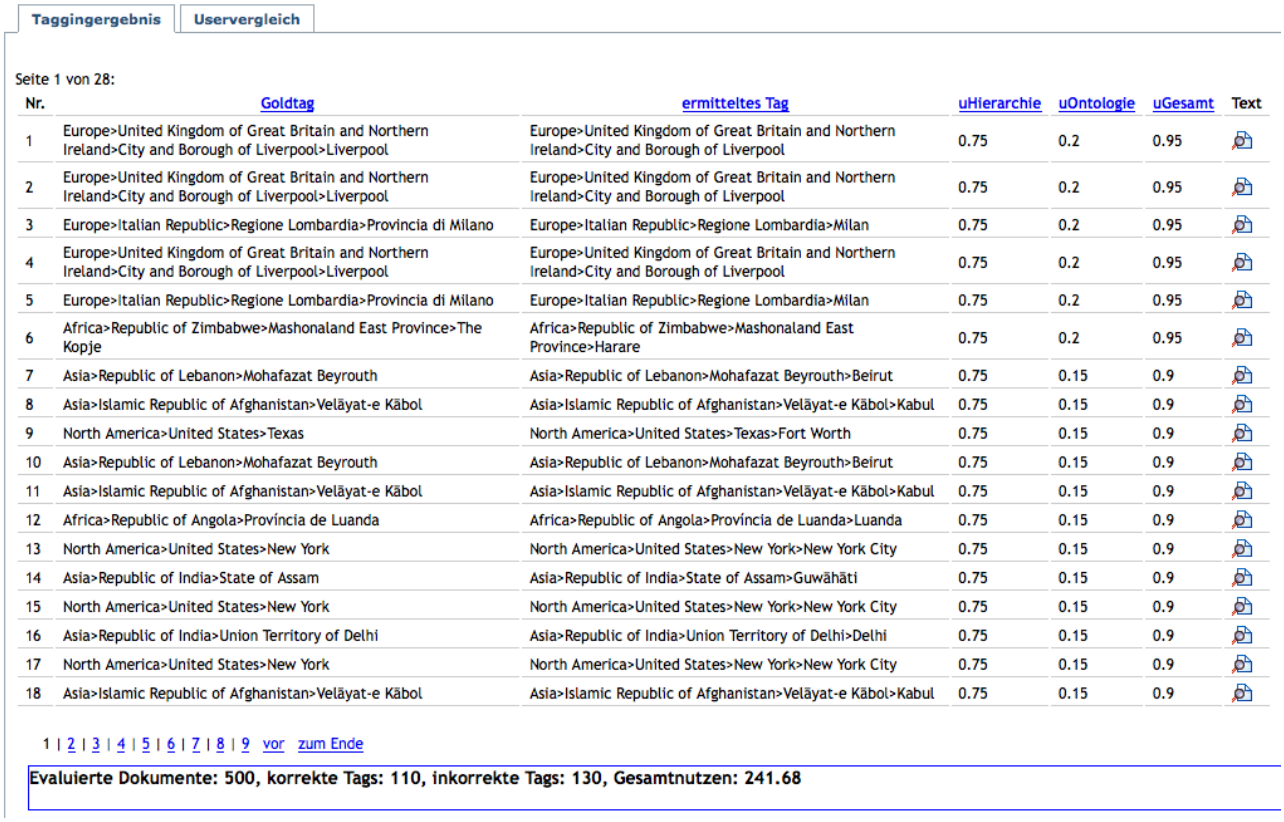
Das Ergebnis mit Gazetteer C500.000 weist Ähnlichkeiten mit jenem mit Gazetteer C100.000 auf. Von einer weiteren Verbesserung, wie sie beim Sprung von Gazetteer C5.000 auf Gazetteer C100.000 der Fall war, ist man jedoch weit entfernt.
Gesamt gesehen konnte die Anzahl an korrekt getaggten Newsartikeln marginal (um 3) gesteigert werden.
Auch der Gesamtnutzen der Usecases "no ontology" sowie "max ontology" wird davon minimal positiv beeinflusst (Verbesserung bei beiden um ca. 1%).

Tabelle 3.10 zeigt das Ergebnis der Evaluierung.

Usecase	korrekt	inkorrekt	Gesamtnutzen	Anzahl Ont-Eval
Wander-Reiseführer	110	130	241,68	22
Europa-Reiseführer	110	130	241,98	206
no Ontology	110	130	238,08	0
max Ontology	110	130	336,09	206

Tabelle 3.10: Geotagging-Ergebnis mit Gazetteer C500.000

Abbildungen 3.11 und 3.12 zeigen die Evaluierungsergebnisse beider Usecases nach Ontologie-Punkten sortiert. Wie erwartet, wird für Usecase eins mehr Nutzen (u_{total}=241,98 vs. u_{total}=241,68) generiert, wobei die Differenz ebenfalls marginal ist (0,30).
Die Minimum-Schranke von 500.000 scheint immer noch zu gering, um deutliche Differenzen zwischen den beiden Usecases zu zeigen. Würde diese im Millionenbereich liegen, würden erstens erheblich mehr Newsartikel korrekt getaggt werden und zweitens mehr Ontologie-Gewichtungen seitens Usecase zwei (”Europa-Reiseführer”) greifen, da nun auch die ermittelten Geo-Tags grober sein würden. Abbildung 3.12 weist nur einen Newsartikel (Nr. 1) auf bei dem beide Tags (Gold-Standard, ermittelt) auf Länderebene sind. Frankreich wird als Nachbarsnachbar von Slowenien (via Italien) angeführt und mit 0,25 Ontologiepunkten bewertet.



Taggingergebnis | Uservergleich

Seite 1 von 28:

Nr.	Goldtag	ermitteltes Tag	uHierarchie	uOntologie	uGesamt	Text
1	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Liverpool>Liverpool	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Liverpool	0.75	0.2	0.95	
2	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Liverpool>Liverpool	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Liverpool	0.75	0.2	0.95	
3	Europe>Italian Republic>Regione Lombardia>Provincia di Milano	Europe>Italian Republic>Regione Lombardia>Milan	0.75	0.2	0.95	
4	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Liverpool>Liverpool	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Liverpool	0.75	0.2	0.95	
5	Europe>Italian Republic>Regione Lombardia>Provincia di Milano	Europe>Italian Republic>Regione Lombardia>Milan	0.75	0.2	0.95	
6	Africa>Republic of Zimbabwe>Mashonaland East Province>The Kopje	Africa>Republic of Zimbabwe>Mashonaland East Province>Harare	0.75	0.2	0.95	
7	Asia>Republic of Lebanon>Mohafazat Beyrouth	Asia>Republic of Lebanon>Mohafazat Beyrouth>Beirut	0.75	0.15	0.9	
8	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol>Kabul	0.75	0.15	0.9	
9	North America>United States>Texas	North America>United States>Texas>Fort Worth	0.75	0.15	0.9	
10	Asia>Republic of Lebanon>Mohafazat Beyrouth	Asia>Republic of Lebanon>Mohafazat Beyrouth>Beirut	0.75	0.15	0.9	
11	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol>Kabul	0.75	0.15	0.9	
12	Africa>Republic of Angola>Província de Luanda	Africa>Republic of Angola>Província de Luanda>Luanda	0.75	0.15	0.9	
13	North America>United States>New York	North America>United States>New York>New York City	0.75	0.15	0.9	
14	Asia>Republic of India>State of Assam	Asia>Republic of India>State of Assam>Guwāhāti	0.75	0.15	0.9	
15	North America>United States>New York	North America>United States>New York>New York City	0.75	0.15	0.9	
16	Asia>Republic of India>Union Territory of Delhi	Asia>Republic of India>Union Territory of Delhi>Delhi	0.75	0.15	0.9	
17	North America>United States>New York	North America>United States>New York>New York City	0.75	0.15	0.9	
18	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol	Asia>Islamic Republic of Afghanistan>Velāyat-e Kābol>Kabul	0.75	0.15	0.9	

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 vor zum Ende

Evaluierte Dokumente: 500, korrekte Tags: 110, inkorrekte Tags: 130, Gesamtnutzen: 241.68

Abbildung 3.11: Evaluierungsergebnisse nach Ontologie-Punkten sortiert ("Wander-Reiseführer"), Gazetteer C500.000

Seite 1 von 28:

Nr.	Goldtag	ermitteltes Tag	uHierarchie	uOntologie	uGesamt	Text
1	Europe>Republic of Slovenia	Europe>Republic of France	0.5	0.25	0.75	
2	South America>Argentine Republic>Distrito Federal>Buenos Aires	South America>Argentine Republic>Provincia de Buenos Aires	0.5	0.105	0.61	
3	Europe>Italian Republic>Regione Lombardia>Provincia di Milano	Europe>Italian Republic>Regione Lombardia>Milan	0.75	0.1	0.85	
4	Europe>Italian Republic>Regione Lombardia>Provincia di Milano	Europe>Italian Republic>Regione Lombardia>Milan	0.75	0.1	0.85	
5	Africa>Republic of Zimbabwe>Mashonaland East Province>The Kopje	Africa>Republic of Zimbabwe>Mashonaland East Province>Harare	0.75	0.1	0.85	
6	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Liverpool>Liverpool	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Liverpool	0.75	0.075	0.83	
7	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Liverpool>Liverpool	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Liverpool	0.75	0.075	0.83	
8	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Liverpool>Liverpool	Europe>United Kingdom of Great Britain and Northern Ireland>City and Borough of Liverpool	0.75	0.075	0.83	
9	Europe>Federal Republic of Germany>Land Niedersachsen>Hameln	Europe>Federal Republic of Germany	0.5	0.06	0.56	
10	Asia>State of Israel	Asia>Palestine>West Bank	0.333333333333	0.0466666666667	0.38	
11	Asia>Republic of the Philippines	Asia>Republic of the Philippines>Province of North Cotabato	0.666666666667	0.0333333333333	0.7	
12	Europe>Italian Republic	Europe>Italian Republic>Regione Calabria	0.666666666667	0.0333333333333	0.7	
13	Asia>Islamic Republic of Pakistan	Asia>Islamic Republic of Pakistan>Federally Administered Tribal Areas	0.666666666667	0.0333333333333	0.7	
14	Europe>Република Македонија	Europe>Република Македонија>Opština Tetovo	0.666666666667	0.0333333333333	0.7	
15	Asia>Republic of Turkey	Asia>Republic of Turkey>İstanbul	0.666666666667	0.0333333333333	0.7	
16	Africa>Federal Republic of Nigeria	Africa>Federal Republic of Nigeria>Niger State	0.666666666667	0.0333333333333	0.7	
17	Europe>United Kingdom of Great Britain and Northern Ireland	Europe>United Kingdom of Great Britain and Northern Ireland>Borough of Trafford	0.666666666667	0.0333333333333	0.7	
18	South America>Bolivarian Republic of Venezuela	South America>Bolivarian Republic of Venezuela>Estado Miranda	0.666666666667	0.0333333333333	0.7	

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 vor zum Ende

Evaluierte Dokumente: 500, korrekte Tags: 110, inkorrekte Tags: 130, Gesamtnutzen: 241.98

Abbildung 3.12: Evaluierungsergebnisse nach Ontologie-Punkten sortiert ("Europa-Reiseführer"), Gazetteer C500.000

Zusammenfassung der Ergebnisse

Tabelle 3.11 stellt die Ergebnisse der beiden Usecases hinsichtlich der verwendeten Gazetteers gegenüber.

Usecase	C5.000	C100.000	C500.000	Verbesserung C5.000 - C500.000
Wander-Reiseführer	151,64	234,95	241,68	159,37%
Europa-Reiseführer	150,53	235,34	241,98	160,75%
no Ontology	148,09	231,45	238,08	160,77%
max Ontology	230,49	331,72	336,39	145,95%

Tabelle 3.11: Gegenüberstellung der Geo-Tagging Ergebnisse beider Usecases

Abbildung 3.13 zeigt eine graphische Darstellung der Ergebnisse:

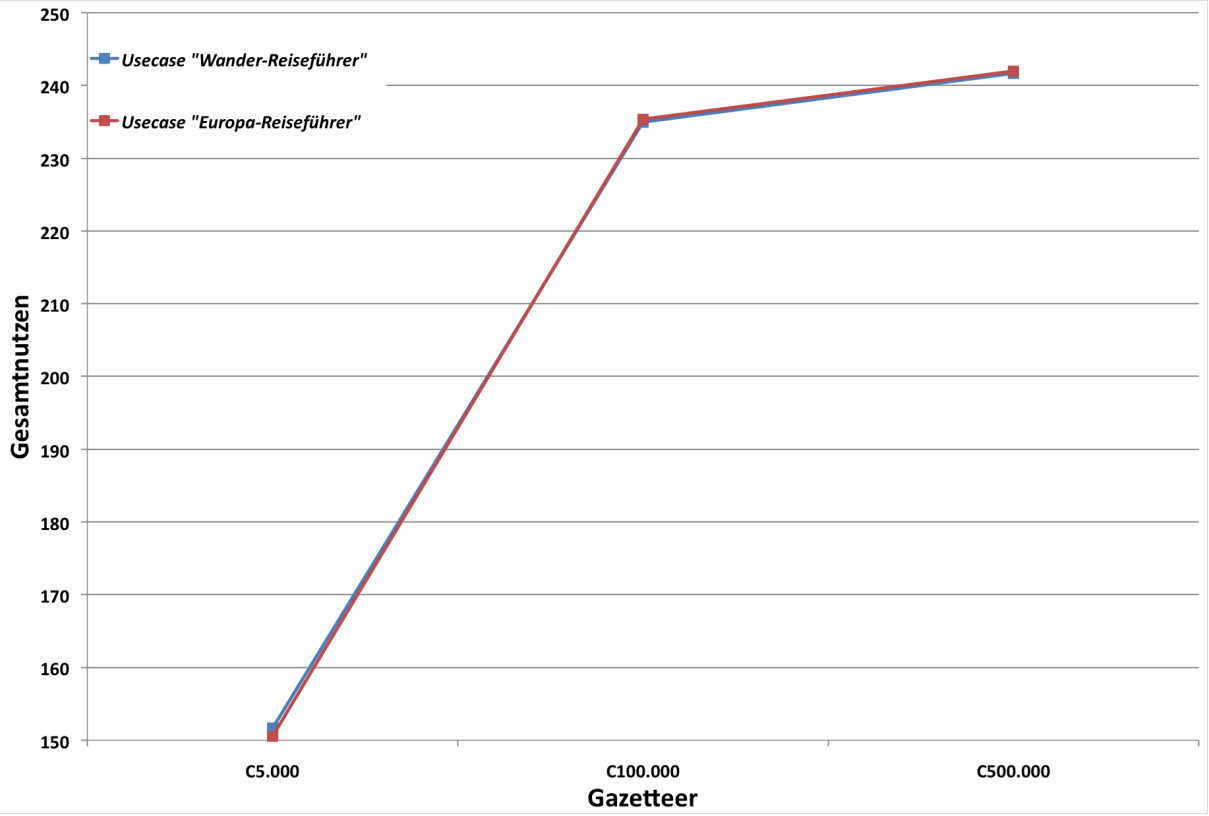


Abbildung 3.13: Graphische Darstellung der Geo-Tagging Ergebnisse

Die Graphik unterstreicht die Empfehlung des Test-Frameworks - nämlich die Verwendung des Gazetteers C500.000 um optimale Ergebnisse für beide

Usecases unter den gegeben Bedingungen zu realisieren. Die Performance verbessert sich durch die Erhöhung der Minimum-Schranke von 5.000 (Gazetteer C5.000) auf 500.000 (Gazetteer C500.000) um 159,27% ("Wander-Reiseführer") beziehungsweise 160,75% ("Europa-Reiseführer").

Man sieht, dass bei den gewählten Gazetteers aufgrund ihren Minimum-Schranken marginale Unterschiede zwischen den beiden Usecases herrschen, obwohl diese aufgrund ihrer Gewichtungen stark unterschiedlich sind. Während der eine Usecase fein-granulare Ergebnisse ("Wander-Reiseführer") präferiert, favorisiert der andere Usecase grob-granulare Ergebnisse ("Europa-Reiseführer"). Der primäre Grund hierfür liegt im tendeziell groben Tagging des Referenz-Geo-Taggers *OpenCalais* (zumeist auf Country-Ebene), der für die Kennzeichnung der Gold-Standard Geo-Tags verantwortlich ist.
Um stärkere Differenzen zwischen den beiden Usecases herauszuarbeiten, hätte die Evaluierung auch Gazetteers mit Minimum-Schranken im Millionen-(Einwohner-)Bereich beinhalten müssen. Dadurch hätte der *geoLyzard* Geotagger öfters auf Country-Ebene getaggt, was wiederum zur Folge hätte, dass sich dies positiv auf den Nutzen von Usecase "Europa-Reiseführer" und zugleich negativ beziehungsweise schwächer positiv auf den Nutzen von Usecase "Wander-Reiseführer" ausgewirkt hätte.
Dies war leider aufgrund technischer Restriktionen nicht möglich. Das verwendete Web-Service des *geoLyard* konnte zu diesem Zeitpunkt nur eine Minimum-Schranke von höchstens 500.000 (Einwohner) verarbeiten.
Abbildung 3.14 soll dennoch anhand der eher unrealistischen Usecases "no Ontology" und "max Ontology" aufzeigen, wie groß das Einflusspotential der Ontologie-Evaluierung ist. Die detaillierten Endergebnisse können Tabelle 3.11 entnommen werden.

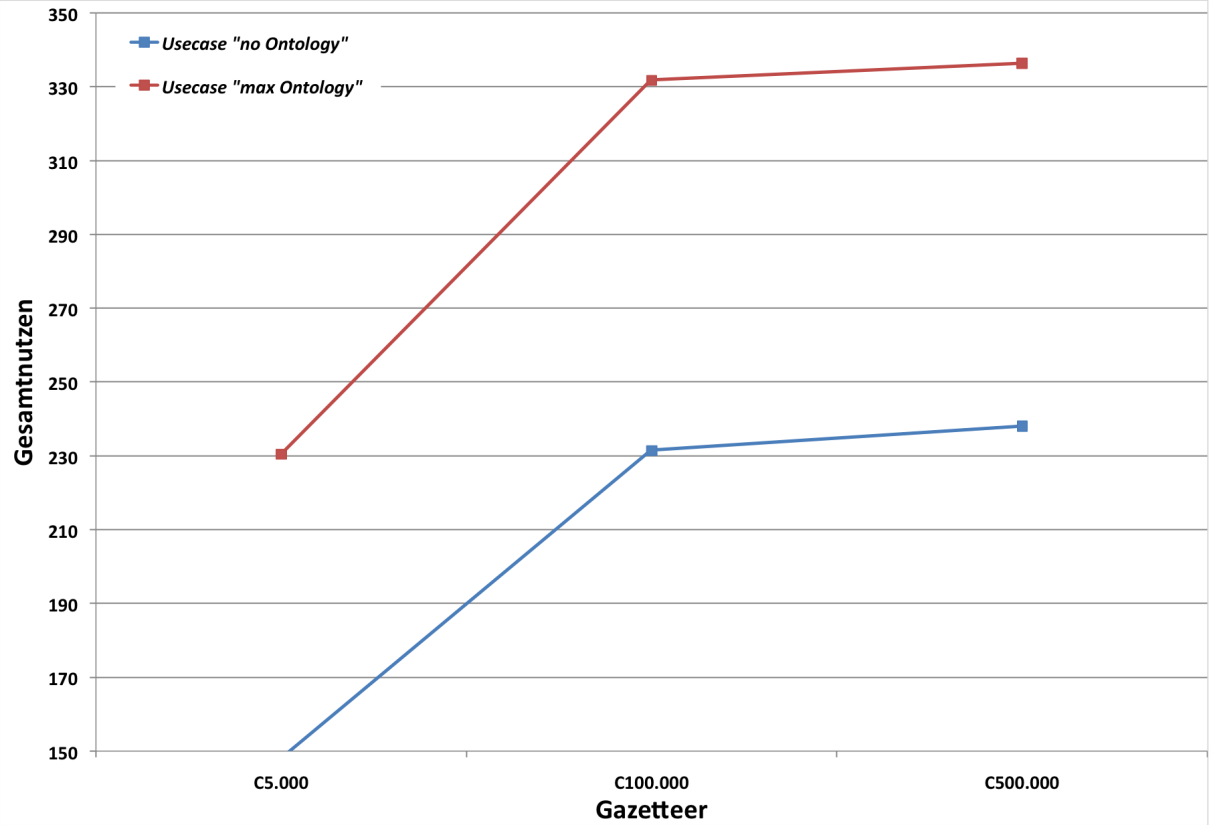


Abbildung 3.14: Gegenüberstellung der Usecases ”no Ontology” und ”max Ontology”

Wie man erkennen kann, kann Usecase ”max Ontology” in etwa den 1,5 fachen Nutzen von Usecase ”no Ontology” generieren, wobei diese starke Differenz auf Basis der benutzerspezifischen Gewichtungen erreicht wurde. Das heißt, das Taggingergebnis ist für beide ident, wird aber aufgrund der Benutzereinstellungen unterschiedlich interpretiert.
Zuletzt wird auch hier die Empfehlung zur Verwendung des Gazetteers C500.000 geraten.

4 Zusammenfassung und Ausblick

Das entwickelte Test-Framework stellt einen ersten Ansatz dar, um Geo-Tags durch Zuhilfenahme des Konzepts des ökonomischen Nutzens und unter der Berücksichtigung von benutzerspezifischen Parametern zu evaluieren. Aufgrund des zeitlichen Limits, welchem diese Arbeit unterliegt, ist es offensichtlich, dass hier noch Optimierungspotential hinsichtlich des Frameworks existiert. Deshalb möchte ich in diesem Abschnitt noch mögliche weitere Verbesserungen adressieren:

Im Bereich der Ontologie-Evaluierung könnte man dem Benutzer durch das **Aufnehmen neuer Ontologie-Vergleiche** mehr Flexibilität bei der Gewichtung geben. Im Moment beschränkt sich dies auf die Evaluierung eines Ontologie-Pfades hinsichtlich *contains*-, *partOf*-, *isNeighbourOf*-, *isNeighbourOfNeighbour*-Beziehungen. Beispielsweise könnte einem Benutzer ein Tagging-Ergebnis mit einem Ort, der Meerzugang hat, mehr Nutzen bringen als ein Ort ohne eines solchen. Keineswegs ist die Evaluierung auf den Vergleich rein geographischer Inhalte beschränkt. Für gewisse Applikationen könnte auch der politische Hintergrund von Relevanz sein. Man könnte so eine weitere semantische Beziehungen von Mitgliedsstaaten der EU, der NATO, der UNO und weiteren Organisationen herstellen und so das Geo-Tagging Ergebnis weiter parametrisieren.
Dies sind nur zwei von vielen Ansätzen um innerhalb des Frameworks mehr Flexibilität zu bieten. Eine Liste der momentan verfügbaren Web-Services des *Geonames*-Gazetteers ist hier verfügbar:
http://www.geonames.org/export/ws-overview.html.

Auch die **Berechnung der Ontologie-Evaluierung** kann noch optimiert werden. Weichselbraun schlägt dazu in [Weichselbraun2009] die Bereinigung dieses Ergebnisses durch einen distanzbasierten Faktor. Es wird ein Verhältnis zweier Distanzen berechnet. Dividend ist die Distanz zwischen dem korrekten Punkt und dem Punkt aus dem ermittelten Geo-Tag. Divisor ist die Distanz zwischen zwei zufällig gewählten Punkten innerhalb einer kreisförmigen Fläche, die der zuletzt korrekt ermittelten Hierarchieebene entspricht ("Austria" bei "Austria/Upper Austria/Linz" und "Austria/Styria/Graz").

Zuletzt darf auch nicht vergessen werden, dass die Veränderung des Tagging-Ergebnisses lediglich aufgrund der Veränderung der Minimum-Schranke (minimale Anzahl an Einwohner pro Geo-Entität) dokumentiert wurde. Dass ein Geotagger, je nach Implementation, **mehrere Konfigurationsparameter** aufweist, wurde bereits eingangs in der Arbeit erwähnt. Mein Betreuer, Dr. Albert Weichselbraun, konnte aufgrund von Erfahrungswerten die Minimum-Schranke als die stärkste Einflussgröße hinsichtlich des Tagging-Ergebnisses identifizieren. Im Falle des hier eingesetzten Geotaggers (*geoLyzard*) wäre zum Beispiel die Art des zu verwendenden Algorithmus ein weiterer Optimierungsparameter.

Literaturverzeichnis

[Ahlers2008] Dirk Ahlers and Susanne Boll. Retrieving address-based locations from the web. In *Proceeding of the 2nd international workshop on Geographic information retrieval*, pages 27–34, Napa Valley, California, USA, 2008. ACM.

[Alani2001] Harith Alani, Christopher B. Jones, and Douglas Tudhope. Voronoi-based region approximation for geographical information retrieval with gazetteers. volume 15, pages 287–306, 2001.

[Allan2005] James Allan, Ben Carterette, and Joshua Lewis. "when will information retrieval be good enough"? In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 433–440, New York, NY, USA, 2005. ACM.

[Amitay2004] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: geotagging web content. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280, New York, NY, USA, 2004. ACM.

[Angel2008] Albert Angel, Chara Lontou, Dieter Pfoser, and Alexandros Efentakis. Qualitative geocoding of persistent web pages. In *Proceedings of the 16th ACM SIGSPATIAL*

international conference on Advances in geographic information systems, pages 1–10, Irvine, California, 2008. ACM.

[Asadi2006] Saeid Asadi, Jiajie Xu, Yuan Shi, Joachim Diederich, and Xiaofang Zhou. Calculation of target locations for web resources. In *Web Information Systems*, pages 277–288. 2006.

[Behr2009] Franz-Josef Behr. Gml-basierte kodierung von geodaten. Technical report, Hochschule für Technik, Stuttgart, 2009. http://www.gis-news.de/papers/gml/gml_paper_part1_3.htm

Abruf am 04.09.2009

[Bertolotto2008] Michela Bertolotto, Cyril Ray, and Xiang Li. *Web and Wireless Geographical Information Systems: 8th International Symposium, W2GIS 2008, Shanghai, China, December 11-12, 2008. Proceedings*. Springer, 1 edition, December 2008.

[Bharti2009] Bharti Softland. CakePHP Developments. http://www.bhartisoftland.com/technologies-skill-sets/cake-php-developments.html

Abruf am 10.11.2009

[Biztech2009] Biztech2 Staff. Unstructured Data Driving CIO Focus To Archival Solns. http://biztech2.in.com/india/news/enterprise-solutions/unstructured-data-driving-cio-focus-to-archival-solns/52832/0

Abruf am 16.9.2009

[Blessing2006] A. Blessing, S. Klatt, D. Nicklas, S. Volz, and H. Schiitze. Language-derived information and context models. In

Pervasive Computing and Communications Workshops, 2006. PerCom Workshops 2006. Fourth Annual IEEE International Conference on, 2006.

[Blessing2008] A. Blessing and H. Schütze. Automatic acquisition of vernacular places. In *Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services*, Linz, Austria, 2008. ACM.

[Blotevogel2002] Blotevogel. Geographie. In *Lexikon der Geographie*, number 2, pages 14–40, Heidelberg/Berlin, 2002.

[CakePHP2009] CakePHP: the rapid development PHP Framework, 2009. http://cakephp.org/

Abruf am 21.12.2009

[Campelo2008] Claudio Elezio Calazans Campelo and Claudio de Souza Baptista. Geographic scope modeling for web documents. In *Proceeding of the 2nd international workshop on Geographic information retrieval*, Napa Valley, California, USA, 2008. ACM.

[Chapelle2006] Oliver Chapelle & Bernhard Schölkopf & Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006.

[Clough2004] Paul Clough and Mark Sanderson. A proposal for comparative evaluation of automatic annotation for geo-referenced documents. In *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*, 2004.

[Densham2003] Ian Densham and James Reid. A Geo-Coding service encompassing a Geo-Parsing tool and integrated digital gazetteer service. *Workshop on the Analysis of Geographic References at the NAACL-HLT 2003 conference*, 2003.

[Dickinger2008] Astrid Dickinger, Arno Scharl, Hermann Stern, Albert Weichselbraun, and Karl Wöber. Acquisition and relevance of geotagged information in tourism. In *Information and Communication Technologies in Tourism 2008*, pages 545–555. 2008.

[Gan2008] Qingqing Gan, Josh Attenberg, Alexander Markowetz, and Torsten Suel. Analysis of geographic queries in a search engine log. In *Proceedings of the first international workshop on Location and the web*, pages 49–56, Beijing, China, 2008. ACM.

[Gartner2009a] Christy Pettey & Holly Stevens. Gartner's 2009 hype cycle special report, 2009. http://www.gartner.com/it/page.jsp?id=1124212

Abruf am 25.9.2009

[Gartner2009b] Christy Pettey & Holly Stevens. Gartner reveals five business intelligence predictions for 2009 and beyond, January 2009. http://www.gartner.com/it/page.jsp?id=856714

Abruf am 25.9.2009

[Geonames2009] Geonames. Geonames - geographical database, 2009. http://www.geonames.org/about.html

Abruf am 16.9.2009

[Glover2002] Eric J Glover, Kostas Tsioutsiouliklis, Steve Lawrence, Steve Lawrence David M Pennock, and Gary W Flake. Using web structure for classifying and describing web pages. pages 562–569, Honolulu, Hawaii, USA, 2002. ACM Press.

[Gravano2003] Luis Gravano, Vasileios Hatzivassiloglou, and Richard Lichtenstein. Categorizing web queries according to geographical locality. In *CIKM '03: Proceedings of the twelfth*

international conference on Information and knowledge management, pages 325–333, New York, NY, USA, 2003. ACM.

[Hersh2000] William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kramer, Lynetta Sacherek, and Daniel Olson. Do batch and user evaluations give the same results? In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 17–24, New York, NY, USA, 2000. ACM.

[Hjelm2002] Johan Hjelm. *Creating Location Services for the Wireless Web*. John Wiley and Sons, February 2002.

[Lang2006] Joel Lang. Named entity recognition. In *Named Entity Recognition*, pages 2–4, 2006.

[Lee2008] Jongwuk Lee and Seung-won Hwang. Ranking with tagging as quality indicators. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 2432–2436, New York, NY, USA, 2008. ACM.

[Leidner2003] J. L Leidner, G. Sinclair, and B. Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 31–38, 2003.

[Leidner2006] Jochen L. Leidner. An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, 30:400–417, 2006.

[Leidner2008] Jochen L. Leidner. *Toponym Resolution in Text*. Universal-Publishers, 2008.

[Li2005] Yaoyong Li, Kalina Bontcheva, and Hamish Cunningham. Svm based learning system for information extraction.

In *In Proceedings of Sheffield Machine Learning Workshop, Lecture Notes in Computer Science*. Springer Verlag, 2005.

[Lyman2003] Peter Lyman and Hal. R. Varian. How much information? 2003, 2003. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf

Abruf am 16.8.2009

[Malczewski1999] Jacek Malczewski. *GIS and multicriteria decision analysis*. John Wiley and Sons, 1999.

[Markowetz2005] Alexander Markowetz, Yen-Yu Chen, Torsten Suel, Xiaohui Long, and Bernhard Seeger. Design and implementation of a geographic search engine. In *WebDB*, pages 19–24, 2005.

[Martins2005] Bruno Martins, Mário J. Silva, and Marcirio Silveira Chaves. Challenges and resources for evaluating geographical ir. In *GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval*, pages 65–69, New York, NY, USA, 2005. ACM.

[Metacarta2009] Metacarta. Metacarta geographic data module - french language. http://metacarta.com/products-data-modules-french.htm

Abruf am 19.9.2009

[Microformats2009] Microformats. Geo microformat specification, 2009. http://microformats.org/wiki/Main_Page

Abruf am 24.9.2009

[Mitchell2008] Tyler Mitchell, Arnulf Christl, and Astrid Emde. *Web-Mapping mit Open Source-GIS-Tools*. O'Reilly Germany, February 2008.

[Moxley2008] Emily Moxley, Jim Kleban, and B. S. Manjunath. Spirittagger: a geo-aware tag suggestion tool mined from flickr. In *Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 24–30, Vancouver, British Columbia, Canada, 2008. ACM.

[Neumann2001] Günter Neumann. Informationsextraktion. In *Computerlinguistik und Sprachtechnologie - Eine Einführung.* Spektrum Akademischer Verlag, Heidelberg, 2001.

[Nielsen1993] Jakob Nielsen and Thomas K. Landauer. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems*, pages 206–213, Amsterdam, The Netherlands, 1993. ACM.

[ODP2010] Open Directory Project. About the open directory project. http://www.dmoz.org/about.html

Abruf am 30.01.2010

[OGC2009] Cliff Kottman and Carl Reed. The OpenGIS Abstract Specification: Topic 5: Features. *Open Geospatial Consortium Inc.*, January 2009. http://portal.opengeospatial.org/files/?artifact_id=29536

Abruf am 10.09.2009

[Pasley2007] Robert C. Pasley, Paul D. Clough, and Mark Sanderson. Geo-tagging for imprecise regions of different sizes. In *GIR '07: Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 77–82, New York, NY, USA, 2007. ACM.

[Pasley2008] Robert Pasley, Paul Clough, Ross S. Purves, and Florian A. Twaroch. Mapping geographic coverage of the web.

In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, pages 1–9, Irvine, California, 2008. ACM.

[Pellegrini2006] *Semantic Web: Wege zur vernetzten Wissensgesellschaft.* Springer, Berlin, May 2006.

[Pospech2009] Thomas Pospech. *GML- Geography Markup Language.* GRIN Verlag, May 2009.

[Rauch2003] Erik Rauch, Michael Bukatin, and Kenneth 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, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[Riekert2002] W.-F. Riekert. Automated retrieval of information in the internet by using thesauri and gazetteers as knowledge sources. volume 8, pages 581–590, 2002.

[Roth2002] Jeanette Roth. Der Stand der Kunst in der Eigennamen-Erkennung - Mit einem Fokus auf Produktnamen-Erkennung. Master's thesis, Universität Zürich, 2002.

[Sanderson2004] M. Sanderson and J. Kohler. Analyzing geographic queries. In *Proceedings Workshop on Geographical Information Retrieval SIGIR*, 2004.

[Scharl2007] Arno Scharl and Klaus Tochtermann. *The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 Are Shaping the Network Society.* Springer, Berlin, 1 edition, May 2007.

[Scherer2009] Dirk Ammelburger und Robert Scherer. *Webentwicklung mit CakePHP.* O'Reilly, 2009.

[Schlieder2006] Christoph Schlieder und Andreas Heinrich und Volker Luedecke und Jens Gräf. Geographisches Information Retrieval. *Datenbank-Spektrum*, 18:48–56, 2006.

[Stock2006] Wolfgang G. Stock. *Information Retrieval*. Oldenbourg Wissenschaftsverlag, October 2006.

[Teitler2008] Benjamin E. Teitler, Michael D. Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Hanan Samet, and Jon Sperling. Newsstand: a new view on news. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, pages 1–10, Irvine, California, 2008. ACM.

[TrendMobi2009] Sven Wiesner. Mit Qype Radar durch den Hamburger Grossstadtdschungel, 2009. http://trendmobi.de/index.php/2009/04/mit-qype-radar-durch-den-hamburger-grosstadtdschungel/

Abruf am 20.9.2009

[Turpin2001] Andrew H. Turpin and William Hersh. Why batch and user evaluations do not give the same results. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 225–231, New York, NY, USA, 2001. ACM.

[Turpin2006] Andrew H. Turpin and Falk Scholer. User performance versus precision measures for simple search tasks. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 11–18, New York, NY, USA, 2006. ACM.

[Weichselbraun2008] Albert Weichselbraun, Arno Scharl, and Hermann Stern. Annotating and visualizing location data in geospa-

tial web applications. In *Proceedings of the first international workshop on Location and the web*, pages 65–68, Beijing, China, 2008. ACM.

[Weichselbraun2009] Albert Weichselbraun. A utility centered approach for evaluating and optimizing geo-tagging. In *First International Conference on Knowledge Discovery and Information Retrieval (KDIR 2009)*, pages 134–139, Madeira, Portugal, October 2009.

[Yanai2009] Keiji Yanai and Bingyu Qiu. Mining cultural differences from a large number of geotagged photos. In *Proceedings of the 18th international conference on World wide web*, pages 1173–1174, Madrid, Spain, 2009. ACM.

[Zhu05] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005. http://pages.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf

Abruf am 20.10.2009

[Zubizarreta2008] Álvaro Zubizarreta, Pablo de la Fuente, José M. Cantera, Mario Arias, Jorge Cabrero, Guido García, César Llamas, and Jesús Vegas. A georeferencing multistage method for locating geographic context in web search. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1485–1486, New York, NY, USA, 2008. ACM.

[derStandard2009] derStandard.at. Augmented Reality wird zum Gemeinschaftserlebnis, August 2009. http://derstandard.at/fs/1250691613547/Layar%20-Augmented-Reality-wird-zum-Gemeinschaftserlebnis

Abruf am 25.9.2009

[eHomeUpgrade2008] eHomeUpgrade. Gartner identifies top ten disruptive technologies for 2008 to 2012, May 2008. http://www.ehomeupgrade.com/2008/05/28/gartner-identifies-top-ten-disruptive-technologies-for-2008-to-2012/

Abruf am 25.9.2009