Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Medieninformatik

Analyse mehrdeutiger Suchanfragen

Bachelorarbeit

Johannes Teschner Matrikelnummer 90076

1. Gutachter: Junior-Prof. Dr. Matthias Hagen

Datum der Abgabe: 10. Juni 2014

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Weimar, den 10. Juni 2014

.....
Johannes Teschner

Zusammenfassung

Diese Arbeit untersucht mehrdeutige Suchanfragen, die bei der Segmentierung von Anfragen bei der Websuche von Interesse sind. Um solche Anfragen untersuchen zu können, werden mit Korpusanalyse des Webis Query Segmentation Corpus von Hagen et al. aus dem Jahr 2012 mehrdeutige Anfragen extrahiert. Im Zuge der Analyse werden Wikipediatitel zum Erkennen bekannter Konzepte in einer Anfrage benutzt. Es wird betrachtet, welcher Wikipediatitel im Fall von Überlappungen mehrerer Wikipediatitel in einer Anfrage von den Annotatoren segmentiert wird. Die Mehrheit segmentiert in Anfragen mit einem kurzen Wikipediatitel, der von einem längeren eingeschlossen wird, den längeren Wikipediatitel. Überlappen sich zwei Wikipediatitel gegenseitig, wird in der Mehrzahl der Fälle der Titel mit der höheren N-Gramm-Häufigkeit gewählt. In der Menge von Suchanfragen, in denen sich Wikipediatitel überlappen, werden mehrdeutige Anfragen gefunden, deren Segmentierungen analysiert werden.
Anschließend wird eine Methode vorgestellt, um inhaltlich ähnliche, aber durch verschiedene Trennstellen unterschiedliche Segmentierungen zu vereinigen. Auf diese Weise lässt sich der Anteil jener Anfragen, bei denen sich die Segmentierer über die Segmentierung uneinig sind, um 7,63 % auf 21,51 % reduzieren. Gleichzeitig wird die Menge der Suchanfragen, bei denen klar eine Segmentierung favorisiert wird, um 10,41 % auf 64,27 % gesteigert.

Inhaltsverzeichnis

Kapitel 1

Einleitung

In dieser Arbeit geht es um mehrdeutige Suchanfragen. Es wird ermittelt, wie häufig mehrdeutige Anfragen vorkommen, wie sie aufgebaut sind und ob sich generelle Merkmale ableiten lassen.

Das Verb `googeln` ist seit 2004 im Duden verzeichnet und ist mittlerweile fast zum Synonym für „im Web suchen“ geworden. Die Nutzung der im World Wide Web (WWW) vorhandenen Informationen ist heute ohne Suchmaschinen für die meisten Menschen nicht mehr denkbar. So hat die Möglichkeit zu suchen nicht nur die Entwicklung des Web maßgeblich geprägt, sondern auch die Art und Weise, wie Anwender mit dem WWW umgehen.

Obwohl Suchmaschinen viel genutzt werden, wissen die wenigsten Benutzer, dass es möglich ist, die Anzahl der Suchergebnisse durch Segmentierungen einzuschränken, um die Genauigkeit der präsentierten Resultate zu erhöhen. Hierbei werden Teile der Anfrage in Anführungszeichen gesetzt, damit der Suchalgorithmus nur Ergebnisse berücksichtigt, bei denen die eingegrenzten Wörter auch inhaltlich ein Konzept bilden und deshalb in Ergebnisdokumenten aufeinanderfolgen. Mehrdeutigkeiten einer Anfrage können durch Segmentieren auf eine Bedeutung reduziert werden. Da die meisten Benutzer aber von der Möglichkeit zu segmentieren keinen Gebrauch machen, gibt es Verfahren, die Anfragen automatisch segmentieren. So können die Ergebnisse auf die wirklich relevanten Treffer begrenzt werden. Um Algorithmen zur automatischen Segmentierung testen zu können, benötigt man Anfragen und die dazu gehörende „richtige“ Segmentierung.

Eine solche Auswahl von Suchanfragen und Segmentierungen, wie auch eine Sammlung von Texten allgemein, wird Korpus genannt. Das Korpus von Bergsma und Wang von 2007, das 500 Anfragen umfasst und vor allem zum Testen von Segmentierungsverfahren verwendet wurde, kann aufgrund seiner geringen Größe als nicht repräsentativ angesehen werden.
Ein größeres Korpus, das Webis Query Segmentation Corpus (Webis-QseC-

10), stellten Hagen et al. 2010 zusammen. Hierin sind über 50 000 Anfragen aus dem 2006 veröffentlichten AOL-Suchlog enthalten. Auf der Crowdsourcingplattform „Amazon Mechanical Turk“ wurden dazu menschliche Segmentierungen gesammelt, sodass jede Anfrage von mindestens 10 unterschiedlichen Menschen segmentiert wurde.

Jedoch gibt es auch nicht immer eine einzige korrekte Art, die Wörter der Suchanfrage zu gruppieren. Das kann dazu führen, dass verschiedene Menschen unterschiedliche Segmentierungen wählen. Zum Beispiel haben im Webis Query Segmentation Corpus nicht alle Segmentierer die gleiche Segmentierung für die Anfrage `graffiti fonts alphabet` gewählt, sodass im Korpus folgende Verteilung zu finden ist.

Segmentierung (S)	Stimmen
`graffiti fonts|alphabet`	5
`graffiti|fonts|alphabet`	3
`graffiti fonts alphabet`	2

In diesem Fall ist klar, dass mit dieser Anfrage ursprünglich nach einem Alphabet, gesetzt in Graffitischriftarten, gesucht wurde. Jedoch gibt es auch Suchanfragen in denen mehrere unterschiedliche semantische Konzepte enthalten sind, was ebenfalls zur Folge haben kann, dass verschiedene Menschen unterschiedliche Einteilungen wählen.

Auf der Suche nach solchen mehrdeutigen Anfragen, bei denen sich Menschen uneinig sind, wie segmentiert werden soll, wird das Webis Query Segmentation Corpus hinsichtlich der Zuverlässigkeit der einzelnen Segmentierer analysiert. Außerdem wird näher beleuchtet, wie sich Suchanfragen, die von Menschen verschieden segmentiert wurden, inhaltlich aber dasselbe Konzept meinen, vereinigen lassen, um die optimale Segmentierung zu erhalten.
In Kapitel 2 werden mit verschiedenen Segmentierungsalgorithmen und Korpora bereits vorhandene Forschungsansätze vorgestellt. Außerdem wird das Webis Query Segmentation Corpus näher untersucht. In Kapitel 3 werden Wikipediatitel als bewährtes Konzept zur Hilfe genommen, um mehrdeutige Suchanfragen zu finden und zu analysieren. In Kapitel 4 wird versucht, die scheinbare Uneinigkeit der verschiedenen Segmentierer, die sich in der Wahl unterschiedlicher Segmentierungen zeigt, durch Unifizierungsmethoden aufzulösen.

Kapitel 2

Anfragen und Segmentierung

Um die Anzahl der Ergebnisse einer Suchanfrage zu begrenzen und so eine bessere Genauigkeit der Treffer zu erhalten, werden beim Segmentieren die einzelnen Wörter einer Anfrage in kleinere Einheiten zerlegt. Inhaltlich zusammengehörende Wörter werden zusammengefasst. So würde man z.B. bei der Sucheingabe `new york city restaurants` den Namen `new york city` in Anführungszeichen setzen und die Anfrage `"new york city" restaurants` abschicken. Aus einer Eingabe $Q = w1\ w2\ w3\ w4\ ...\ w_n$ ergeben sich 2^{n-1} mögliche Segmentierungen, da $n-1$ Positionen potenzielle Trennstellen sind. Allerdings benutzen gerade einmal 1,12 % der Menschen Anführungszeichen oder Operatoren zur genaueren Suche [WM07]. Deshalb wurden verschiedene Verfahren entwickelt, um vom Benutzer nicht segmentierte Anfragen automatisch in Segmente zu zerlegen.

2.1 Segmentierungsverfahren

Es gibt zwei Kriterien für die Bewertung von Segmentierungsverfahren: die Ähnlichkeit zur menschlichen Segmentierung und die Retrieval Performance. Anhand von Referenzsegmentierungen kann getestet werden, wie menschenähnlich die Ergebnisse eines Segmentierungsalgorithmus sind. Wenn eine Segmentierungsmethode gleiche oder ähnliche Segmentierungen wie Menschen produziert, heißt das nicht aber unbedingt, dass damit auch die Qualität der Suchergebnisse hoch ist, weil Menschen nicht immer optimal segmentieren. Deshalb werden Segmentierungsalgorithmen auch anhand der Ergebnisdokumente, die mit der maschinell segmentierten Anfrage gefunden werden, bewertet. Diese sogenannte Retrieval Performance hat eine größere Bedeutung, da es bei einer Suchmaschine vorallem darum geht, möglichst gute Ergebnisse zu finden.

Mutual Information

Die erste Methode zur Segmentierung von Anfragen, basierend auf Anfragen aus einem Suchlog, wurde 2003 von Risvik et al. vorgestellt [RMB03]. Dabei wird zuerst für alle möglichen Segmentierungen der Anfrage das Produkt aus der absoluten Häufigkeit und der „Mutual Information“ (Transinformation), „Connexity“ genannt, berechnet. Die Häufigkeit bezeichnet dabei, wie oft das Segment in einem Anfragenlog vorkommt. Die Transinformation gibt an, wie wahrscheinlich die Wörter in einem Dokument zusammen zu finden sind. Anschließend wird so segmentiert, dass immer ein maximaler Score-Wert, der sich aus der Summe der Segmente mit mehr als einem Wort ergibt, gewählt wird [RMB03].
Bei Jones et al. muss der Transinformationswert benachbarter Wörter über einem bestimmten Schwellwert liegen, damit diese Wörter als zu segmentierendes Konzept erkannt werden [JRMG06].
Huang et al. verwenden ebenfalls Mutual Information [HGM$^+$10]. Um Anfragen mit vier oder mehr Wörtern besser segmentieren zu können, zerlegen sie jede Anfrage in einen Binärbaum. Dabei wird der Baum von oben nach unten konstruiert. Solange bei einem Knoten der niedrigste Transinformationswert zwischen zwei Segmenten über einem bestimmten Schwellwert liegt, bilden diese zwei Segmente neue Kindknoten. Zum Ermitteln des Transinformationswertes werden N-Gramm-Häufigkeiten verwendet. Wenn ein Knoten schon die kleinstmögliche Länge erreicht hat, wird nicht weiter aufgeteilt.

Segmentierungsalgorithmen, die auf Transinformation alleine setzen, erzielen verhältnismäßig niedrige Performance-Werte und werden deshalb oft als Baseline-Algorithmen zum Vergleich mit anderen neuen Ansätzen verwendet.

Überwachtes Lernen

Bergsma und Wang wählen überwachtes Lernen für ihren Ansatz. In ihrem Algorithmus trainieren sie eine Support Vector Machine auf einem umfangreichen Trainingset [BW07]. Bendersky et al. setzen auf die Kombination eines Markov Random Fields mit einem Segmentierer, der durch überwachtes Lernen trainiert wird [BCS09].

Der Nachteil an den auf überwachtem Lernen basierenden Methoden ist, dass eine große Menge an Trainingsdaten, von Menschen segmentierte Anfragen, vorhanden sein muss. Der Algorithmus lernt menschenähnlich zu segmentieren, was nicht automatisch zur besten Auswahl von Suchergebnissen führen muss.

N-Gramm-Häufigkeiten

Tan und Peng verwenden in ihrem Ansatz einen Expectation-Maximization-Algorithmus [TP08]. Die Anfragen werden dabei mit N-Gramm-Häufigkeiten durch Hinzunahme von Wikipediatiteln segmentiert. Tan und Peng benutzen für diese Methode, die auf unüberwachtem Lernen basiert, 1 % der von der Suchmaschine Yahoo! indizierten Dokumente im Web.
Hagen et al. verfolgen einen naiven Ansatz [HPSB10]. Sie berechnen für jede mögliche Segmentierung einen Score aus der Summe von N-Gramm-Häufigkeiten, nach Anfragenlänge gewichtet. Mit dieser Methode erreichen sie zu vorhandenen Algorithmen vergleichbare Resultate. Sie konnten ihren Algorithmus noch verbessern, in dem ein Segment, das auch ein Wikipediatitel ist, ein höheres Gewicht erhält [HPSB11]. Noch bessere Ergebnisse konnten Hagen et al. mit einem hybriden Algorithmus erzielen, bei dem Strict Noun Phrases (SNP), Anfragen, welche ausschließlich aus Substantiven, Artikeln, Adjektiven und Zahlen bestehen, separat behandelt werden [HPBS12]. Da Benutzer bei SNP-Anfragen öfter mehr Wörter zu Segmenten zusammenfassen als bei Nicht-SNP-Anfragen, werden bei letzteren nur Wikipediatitel als Segmente gewählt. Bei SNP-Anfragen dagegen müssten alle Begriffe segmentiert werden, die mehr als 40 mal im Web auftauchen, um eine menschenähnliche Segmentierung zu erhalten. Die Qualität der Suchergebnisse lässt sich noch durch das Nicht-Segmentieren von SNP-Anfragen steigern, sodass die Suchergebnisqualität anderer Ansätze übertroffen wird. Bei ungünstig gewählten Segmentierungen wird die angezeigte Suchergebnisauswahl im Vergleich zur nicht segmentierten Anfrage verschlechtert, weshalb im Zweifel auf das Segmentieren verzichtet werden sollte [HPBS12].

Conditional Random Fields

Guo et al. benutzen 10 000 Anfragen aus einem Suchanfragenlog, um damit die Transinformation aller 2-Wort-Paare zu ermitteln, die anschließend zur Bestimmung der Trennstellen einzelner Segmente verwendet wird [GXLC08]. Dies passiert im Rahmen der Neuformulierung von Anfragen. Parallel wird noch die Korrektur von Rechtschreibfehlern, das Trennen und Aufspalten von Wörtern unter der Verwendung von Conditional Random Fields (CRFs) durchgeführt.
Yu und Shi verwenden für das Zuordnen von Labels für die einzelnen Schlüsselwörter ebenfalls Conditional Random Fields [YS09]. Als Trainingsdaten werden Einträge aus einer Datenbank benutzt. Anschließend werden die Anfragen mit Labels versehen, die neben einer Zuordnung in eine Kategorie auch beinhalten, ob das Wort am Anfang oder am Ende eines Begriffs steht. Anhand

der zugeordneten Labels wird dann mit einem Score-Maximierungsalgorithmus segmentiert.
Ein ähnlicher Ansatz wird von Kiseleva et al. verfolgt, jedoch versuchen sie schon vor dem Zuordnen der Labels durch CRFs die Schlüsselwörter anhand von Klickdaten einer Kategorie zuzuordnen [KGA+10]. Beide Methoden beziehen sich auf die Suche in relationalen Datenbanken, deshalb sind sie nicht direkt auf die Websuche übertragbar.

Benutzung von angeklickten Ergebnissen

Li et al. verwenden Paare von Anfragen und angeklickten Ergebnisdokumenten [LHZW11]. Dabei werden die Segmente auf Basis der Wort-N-Gramme, welche im angeklickten Dokument vorkommen, gewählt.
Zhang et al. gehen von einem vorhandenen Segmentierungsalgorithmus aus, der eine Rangliste von Segmentierungen ausgibt und nehmen an, dass Anfragen mit gleichen Suchergebnissen inhaltlich ähnlich sein müssen [ZCL+13]. Sie verwenden Anfragen und dazugehörige angeklickte Ergebisse einer Suchmaschine, um ihren Algorithmus, basierend auf einer Support Vector Machine, zu trainieren. Es wird angenommen, dass richtige Segmente öfter in angeklickten Ergebnisdokumenten auftauchen als falsche. Es wird die Segmentierung ausgewählt, bei der die meisten Segmente mit denen der anderen inhaltlich gleichen Anfragen übereinstimmen. Suchanfragen gelten als inhaltlich gleich, wenn für sie die gleichen Suchresultate ausgegeben werden [ZCL+13].

Verwendung der Ergebnisreihenfolge

Ding et al. verwenden die Topergebnisse einer Suchmaschine zum Segmentieren [DDQ+13]. Dazu wählten sie aus einem Anfragenlog 12 396 Anfragen aus, die inhaltlich in bestimmte Bereiche gehören. Anschließend wurden die Anfragen manuell segmentiert und jedes Segment einem Schema (Label) zugeordnet. Der Algorithmus segmentiert die Anfrage anhand von gewichteten Tokens. Die Gewichte der Tokens werden anhand der Position in der Ergebnisliste für die unsegmentierte Anfrage berechnet.

Der Ansatz von Wu et al. ermittelt die beste Segmentierung in zwei Schritten [WHLC13]. Zuerst wird mit der Wikipedia-Normaliserung von Hagen et al. segmentiert [HPSB11]. Dabei bekommen Wikipediatitel ein Gewicht, das sich aus der maximalen Häufigkeit, die sich aus den 2-Grammen des Wikipediatitels zusammensetzt, addiert mit der Segmentlänge, ergibt. Das Ranking der Suchergebnisse dieser Segmentierung wird anschließend für überwachtes Lernen benutzt, um die beste Segmentierung zu finden.

Pseudo Relevance Feedback

Bendersky et al. stellen eine Methode vor, die Pseudo Relevance Feedback (PRF) zur Segmentierung von Anfragen verwendet [BCS10]. Eine bestimmte Anzahl der obersten Ergebnisse werden beim PRF herangezogen, um daraus Informationen zu gewinnen, die für eine Neuformulierung der Anfrage benutzt werden. Bei Bendersky et al. finden N-Gramm-Häufigkeiten Verwendung, um so, anhand der größten Wahrscheinlichkeit von Trennstellen und unter Einbeziehung von Pseudo Relevance Feedback von vorher eingeholten Ergebnislisten für die Suchanfrage, zu segmentieren. So fließt der Rang einer vorhergegangenen Suche als Relevanz mit in die Auswahl der Segmentierung mit ein. Der Nachteil an dieser Vorgehensweise ist, dass schon eine Suchanfrage gestellt werden muss, bevor überhaupt segmentiert wurde, was sehr zeitaufwendig ist [BCS10, BCS11].

Andere Verfahren

Zhang et al. verwenden die Auftrittshäufigkeiten von allen möglichen Kombinationen zweier Wörter einer Anfrage im Web, um daraus eine Matrix zu konstruieren [ZSH+09]. Mit verschiedenen Matrixtransformationen berechnen sie so die Zusammengehörigkeit von Wörtern einer Anfrage, um die Segmentierung zu ermitteln.
Brenes et al. stellen eine Methode vor, die mithilfe von Wortsequenzen, die ein Schlüsselwort (Snippets) enthalten, N-Gramme zählt und so anhand von N-Gramm-Häufigkeiten, solange sie über einem bestimmten Schwellwert liegen, segmentiert [BGAG10]. Dabei werden die Snippets aus Suchergebnissen, die mit wenigen Worten das Ergebnisdokument beschreiben, gewonnen. Für Fälle, in denen noch keine passenden Snippets vorliegen, muss so ersteinmal eine zusätzliche Anfrage durchgeführt werden, was den zeitlichen Aufwand deutlich erhöht.
Mishra et al. stellen 2011 eine Methode vor, die die Häufigkeit von Mehrwortausdrücken in einem Anfragenlog ermittelt und daraus einen Score berechnet, sodass die Anfrage so in Segmente zerlegt wird, dass ein maximaler Wert aus der Summe der einzelnen Segmentwerte erreicht wird [MSRG+11].

2.2 Korpora

Eine Zusammenstellung von Texten, hier Suchanfragen, wird Korpus genannt. Für Segmentierungsverfahren, die mit überwachtem Lernen trainiert werden und zur Evaluierung von maschinell erzeugten Segmentierungen werden Korpora mit von Menschen segmentierten Anfragen benötigt. Da Menschen ver-

schieden segmentieren, reicht es nicht aus, nur eine Person pro Anfrage segmentieren zu lassen. Das Bearbeiten von mehreren menschlichen Segmentierern führt dazu, dass sich deren „Stimmen“ auf unterschiedliche Segmentierungen verteilen.

2.2.1 Andere Korpora

Das Webis Query Segmentation Corpus ist aktuell das größte Korpus im Bereich der Suchanfragensegmentierung. Daneben gibt es noch kleinere, ältere Korpora, die zuerst vorgestellt werden sollen.

Bergsma-Wang-Korpus

Das 2007 veröffentlichte Korpus von Bergsma und Wang war die erste größere verfügbare Zusammenstellung von Suchanfragen, ihrer Segmentierung und der Domain des ausgewählten Ergebnisses. Dafür wurden 500 Anfragen mit einer Länge von vier oder mehr Wörtern aus dem 2006 veröffentlichten AOL-Suchlog verwendet. Es wurden ausschließlich solche Suchanfragen berücksichtigt, für die vom Benutzer auch eines der aufgelisteten Ergebnisse ausgewählt wurde, weil sie durch weniger vorhandene Rechtschreibfehler besser zu verwenden waren. Außerdem musste die Suchanfrage aus Artikeln, Substantiven und Adjektiven bestehen.
Drei verschiedene Menschen segmentierten die 500 Anfragen und hatten dabei Zugriff auf die Domain des ausgewählten Resultates, welches vom ursprünglichen Suchanfragensteller angeklickt worden war [BW07].

Korpus von Bendersky et al.

Das Korpus von Bendersky et al. von 2010 umfasst 250 Suchanfragen aus dem Anfragenlog einer Suchmaschine, bestehend aus 96 Aussagesätzen, 93 Fragen und 61 Anfragen nach Begriffen ohne grammatischen Zusammenhang [BCS10]. Es existieren darin Anfragen mit jeweils ein bis zwölf Wörtern. Für jede Anfrage gibt es genau eine, maschinell erzeugte, Segmentierung. Davon bestehen 37 Anfragen nur aus ein oder zwei Wörtern und sind so für Suchanfragensegmentierung gar nicht oder nur sehr begrenzt sinnvoll nutzbar.

Korpus von Roy et al.

Die Zusammenstellung von Roy et al. umfasst 500 zufällig aus einem Suchanfragenlog von Bing ausgewählte Anfragen von 2010. Dabei wurden nur Anfragen berücksichtigt, die mehr als fünf Mal auftauchen, da so Schreibfehler minimiert werden konnten und die aus mindestens fünf und höchstens acht Wörtern

bestehen. Drei Menschen segmentierten alle Anfragen jeweils unabhängig voneinander [SRGCL12].

2.2.2 Das Webis Query Segmentation Corpus

Das Webis Query Segmentation Corpus (Webis-QSeC-10), das 53 432 Anfragen enthält, wurde 2012 von Hagen et al. veröffentlicht. Alle bis dahin vorhandenen Korpora konnten nicht als repräsentativ angesehen werden, da sie eine zu kleine Anzahl Suchanfragen enthielten und nur von wenigen Menschen segmentiert worden waren. Für das Webis-QSeC-10 wurde zufällig aus dem im Jahr 2006 veröffentlichten AOL Search Log ausgewählt, nachdem unpassende Anfragen (bestehend aus URLs, von Bots etc.) herausgefiltert worden waren. Es fanden ausschließlich Anfragen der Länge drei bis zehn Verwendung, die, nachdem die Rechtschreibung korrigiert worden war, dann jeweils von mindestens zehn verschiedenen Arbeitern bei der Crowdsourcingplattform „Amazon Mechanical Turk“ bearbeitet wurden. Mit Testanfragen wurde überprüft, dass die Arbeiter nicht willkürlich segmentierten. War dies der Fall, dann wurden die Ergebnisse nicht gewertet [HPSB11].

Erstellung des Webis Query Segmentation Corpus

Zunächst wurde der AOL-Suchlog bereinigt. Dazu wurden Anfragen entfernt, die URLs oder nicht-alphanumerische Zeichen außer Bindestriche und Apostrophe enthielten. Um automatisch erzeugte Anfragen auszuschließen, wurden Daten der Benutzer, die mehr als 10 000 Anfragen gestellt hatten oder durchschnittlich weniger als eine Sekunde zwischen zwei gestellten Anfragen verstreichen ließen, nicht berücksichtigt. Außerdem fanden Anfragen von Benutzern keine Verwendung, deren durchschnittliche Anfragenlänge mehr als 100 Buchstaben betrug. Es wurden ausschließlich Suchanfragen mit mindestens drei und maximal zehn Wörtern einbezogen. Aus den übrigen 6 027 600 Anfragen wurden zufällig 55 555 ausgewählt. Anschließend entfernten Hagen et al. ethisch fragwürdige Anfragen sowie solche, die nicht englischer Sprache waren und führten Rechtschreibkorrekturen durch. Danach blieben 53 432 Anfragen übrig.

Diese verbleibenden Suchanfragen wurden gebündelt von Arbeitern der Crowdsourcing-Plattform „Amazon Mechanical Turk“ segmentiert. Vier Anfragen mussten zusammen mit einer Testanfrage als ein „Bündel“ bearbeitet werden. Entsprach die Testanfrage nicht den gültigen Lösungsmöglichkeiten, wurden alle fünf Anfragen abgelehnt und der Arbeiter erhielt keine Bezahlung. So wurde sichergestellt, dass nicht wahllos segmentiert werden konnte. Die angenommenen Segmentierungen wurden ins Webis Query Segmentation

Corpus aufgenommen, sodass für jede Anfrage zehn Segmentierungen von zehn verschiedenen Menschen enthalten sind.

Anzahl bearbeiteter Anfragen und Qualität der Segmentierung

Je mehr Anfragen von einer Person bearbeitet wurden, desto zuverlässiger funktionierte das Segmentieren. Die Anfragen wurden in Blöcken zu je fünf Anfragen und einer Testanfrage bearbeitet. Weil ein Block nur angenommen wurde, wenn die Testanfrage eine gültige Segmentierung aufwies, ergibt sich für die 216 Arbeiter, die jeweils nur fünf Anfragen segmentieren, eine Akzeptanzrate von 100 % (vgl. Abb. 2.2). Insgesamt mussten jedoch 23 % aller Blöcke von Arbeitern, die nur fünf Anfragen bearbeiteten, abgelehnt werden.

Diskussion der Segmentierungen

Es kann unterschiedliche Gründe haben, dass den Menschen, die als Arbeiter bei der Webplattform Amazon Mechanical Turk die Anfragesegmentierung des Webis Query Segmentation Corpus vorgenommen haben, Fehler unterlaufen sind. Eine Originalanfrage könnte schlecht formuliert gewesen sein, sodass ein Arbeiter nicht durch Lesen der Anfrage verstehen konnte, wonach ursprünglich gesucht werden sollte. Das Resultat daraus könnte eine falsche oder gar keine Segmentierung sein.
Außerdem ist zu berücksichtigen, dass die Segmentierer an dem Experiment teilnahmen, um Geld zu verdienen, nicht allein um die Forschung zu unterstützen. Es kann passiert sein, dass Arbeiter aus Gründen der Zeitersparnis, ohne nachzudenken, zufällig gewählte Segmentierungen wie zum Beispiel `general|motors email` abschickten.
Die fehlende Kenntnis des Konzepts hinter dem Suchbegriff könnte ein weiterer Grund für fehlerhafte Segmentierungen wie `john maynard|keynes` sein. Auch nicht ausgeschlossen werden kann, dass Segmentierer versehentlich die noch nicht vollständig bearbeitete Anfrage abschickten. Beispielsweise gibt es für die Anfrage `here we go lyrics` neun Segmentierungen der Form `here we go|lyrics`, aber auch eine, bei der gar nicht segmentiert wurde. Da der Arbeiter aber nur 19 Sekunden (Durchschnitt aller Segmentierer: 59,27 s) Bearbeitungszeit für den Anfragenblock benötigte, ist da von auszugehen, dass die Aufgabe versehentlich zu früh abgeschickt wurde. Die genannten Beispiele sind aber im Korpus vorhanden und zeigen, dass es nicht möglich war, Segmentierungen schlecht arbeitender Segmentierer gänzlich herauszufiltern.

Kritik

Ramanath et al. stellen in Frage, dass Crowdsourcing zur Gewinnung von Referenzsegmentierungen geeignet ist [RCBR13]. Besonders bei der Segmentierung von künstlich generierten Anfragen durch Crowdsourcing zeigte sich, dass die Segmentierer hauptsächlich zwei oder drei Segmente erzeugten. Außerdem fanden Ramanath et al. heraus, dass die Segmente häufig eine ähnliche Länge aufwiesen und ausgeglichene Segmentstrukturen bevorzugen. Es ist jedoch zu bezweifeln, dass die Ergebnisse von Ramanath et al. repräsentativ sind. Zum einen waren die Testanfragen, die am häufigsten in Zwei- oder Dreiwortsegmente zerlegt wurden, künstlich generiert und enthielten selten inhaltlich sinnvolle Konzepte. Daher ist davon auszugehen, dass die Arbeiter beliebig und schnell segmentierten, um weniger Zeit auf die einzelne Aufgabe zu verwenden und so mehr Geld zu verdienen. Zum anderen fand keine Überprüfung der abgeschickten Lösungen statt, sodass die Segmentierer frei beliebige Segmentierungen abschicken konnten, ohne befürchten zu müssen, nicht bezahlt zu werden.

Eigenschaften des Korpus

Das Webis Query Segmentation Corpus wurde erstellt, um Ergebnisse von Algorithmen zur maschinellen Anfragesegmentierung mit den Segmentierungen von Menschen zu vergleichen. So kann überprüft werden, wie menschenähnlich ein Algorithmus segmentiert. Von Anfragen, bei denen sich alle Segmentierer einig sind, bis hin zu Anfragen, bei denen alle Arbeiter unterschiedlich segmentierten, sind alle Kombinationen vorhanden. In letzterem Fall ist keine Segmentierung häufiger als eine andere. Es ist nicht sicher, wie Menschen eine solche Anfrage segmentieren würden. Wenn es nur eine einzige Segmentierung für eine Anfrage gibt, kann ziemlich sicher davon ausgegangen werden, dass das die Segmentierung ist, die Menschen für richtig halten. Deshalb werden die Anfragen in verschiedene Kategorien aufgeteilt. Außerdem muss die Anfragenlänge mit betrachtet werden, denn kürzere Anfragen bestehen meist nur aus Substantiven, Artikeln, Adjektiven und Zahlen, was bei längeren selten ist.

Einteilung in Kategorien

Da manche der 53 432 Anfragen im Webis Query Segmentation Corpus von mehr als zehn Menschen segmentiert wurden, musste ein Teil davon auf eine Stimmverteilung von insgesamt zehn „Stimmen“ normalisiert werden. So können die Segmentierungsergebnisse für jede Anfrage in die Kategorien „sicher“, „mittelsicher“ und „unsicher“ (vgl. Tab. A.1 auf Seite 39 u. ff.) eingeteilt werden. Während die menschlichen Segmentierer für Anfragen der Kategorie

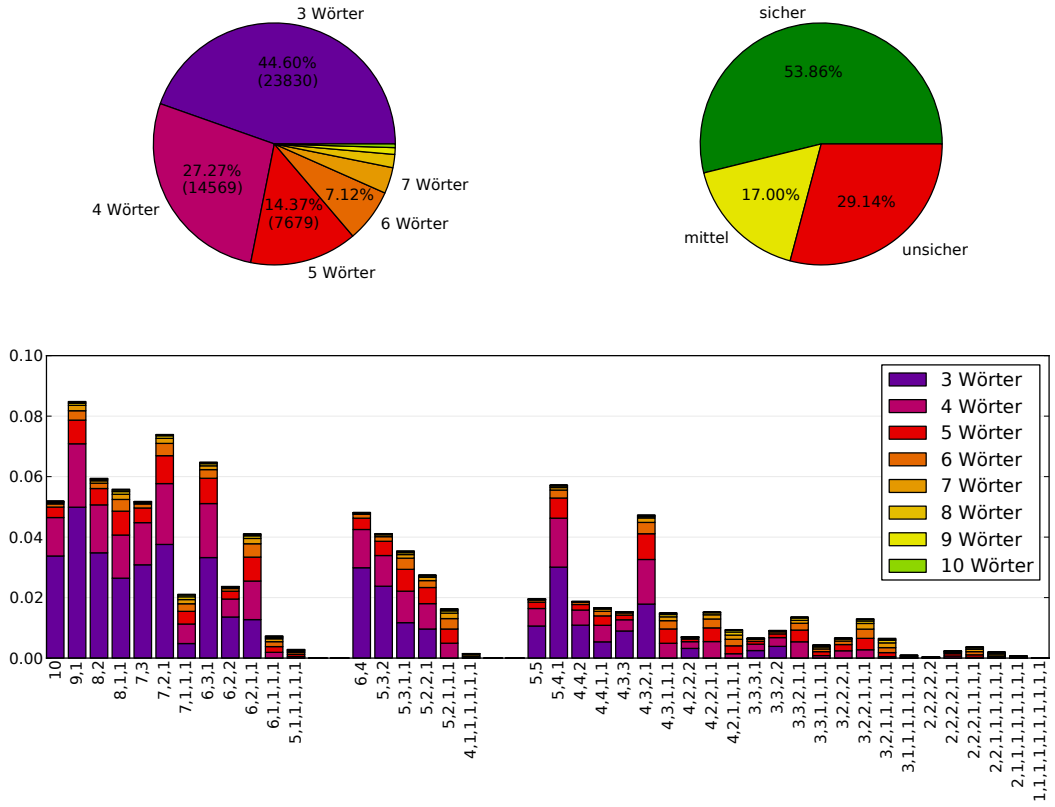

Abbildung 2.1: Verteilung der Segmentierungen im Webis Query Segmentation Corpus. Das Histogramm ist eingeteilt in „sichere“ Stimmverteilungen (links), „unsichere“ Stimmverteilungen (rechts) und die Stimmverteilungen dazwischen (mitte).

„sicher“ ziemlich klar eine Segmentierung bevorzugen, zeigt sich für Anfragen der Kategorie „unsicher“ die Uneinigkeit, wie zu segmentieren ist (vgl. 2.1).

Anfragenlängen

Anfragen mit relativ kurzer Länge machen einen Großteil der sicheren Anfragen aus, was auch damit zusammenhängt, dass bei kürzeren Anfragen durch weniger potenzielle Trennstellen auch insgesamt weniger unterschiedliche Segmentierungen möglich sind. Bei Anfragen der Kategorie „sicher“ kann davon ausgegangen werden, dass die Segmentierung, die am häufigsten gewählt wurde, von Menschen als richtige Art angesehen wird, die Suchwörter zu gruppieren.

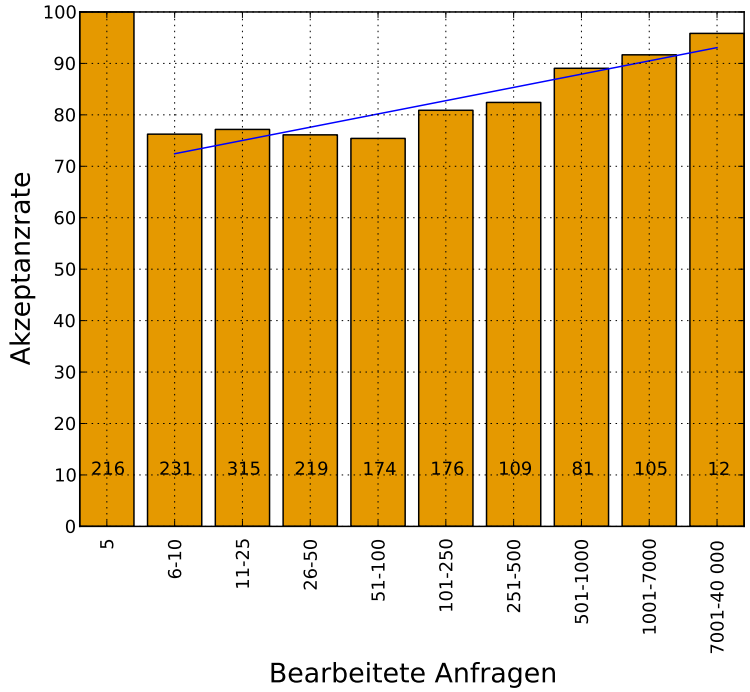

Abbildung 2.2: Die durchschnittliche Akzeptanzrate der Arbeiter, deren Segmentierungen ins Webis Query Segmentation Corpus aufgenommen wurden. In den Balken ist die Anzahl der Arbeiter abzulesen, die eine bestimme Anzahl von Anfragen segmentierten. Es zeigt sich, dass mit zunehmender Zahl bearbeiteter Anfragen die Akzeptanzrate steigt.

In der Kategorie „unsicher“ gibt es mehr längere Anfragen, weil hierbei auch mehr unterschiedliche Segmentierungsmöglichkeiten bestehen.

Hagen et al. kommen zu dem Schluss, dass es für eine gute Retrieval Performance nicht nötig ist, SNP-Anfragen, die sich ausschließlich aus Substantiven, Artikeln, Adjektiven und Zahlen zusammensetzen, zu segmentieren [HPBS12].

Auftreten von Strict Noun Phrases (SNP)

Kürzere Anfragen bestehen häufiger aus Strict Noun Phrases. So sind 67 % der 3-Wort-Anfragen SNPs, bei 4-Wort-Anfragen sind es immer noch 45,71 %. Längere Anfragen dagegen enthalten öfter Wortgruppen oder sogar vollständige Sätze. So finden sich unter den Anfragen mit sechs Wörtern nur noch 12,78 % SNP-Anfragen, bei den 7-Wort-Anfragen sogar nur noch 7,73 %. Hierbei werden von Benutzern seltener Wörter zu einem Segment zusammengefasst als bei Nicht-SNP-Anfragen [HPBS12].

Kapitel 3

Mehrdeutige Suchanfragen

Eine Annahme bei der Suche nach mehrdeutigen Suchanfragen ist, dass diese vor allem in Anfragen zu finden sind, bei denen sich die menschlichen Segmentierer uneinig sind und die deshalb als „unsicher“ eingestuft werden. Im Folgenden werden Anfragen mithilfe von Wikipediatiteln analysiert. Dabei werden Suchanfragen mit überlappenden Wikipediatiteln untersucht. Aus der Menge dieser Suchanfragen werden schließlich mehrdeutige Anfragen gewonnen und genauer betrachtet.

3.1 Wikipediatitel in Suchanfragen

Wikipedia ist die größte Enzyklopädie im World Wide Web. Darin enthalten sind Informationen aller möglichen Themenbereiche, die einem weiteren Kreis von Benutzern bekannt sind. Die Inhalte werden schnell aktualisiert, sodass sie auch die neuesten Trends abbilden. Deshalb eignen sich die Titel der Artikel, zumindest teilweise, um automatisch relevante Konzepte in Suchanfragen zu erkennen [TP08].

Nachteile

Wikipedia enthält teilweise auch Titel von z.B. relativ unbekannten Liedern oder Namen von Episoden einer Fernsehserie, die selten in Suchanfragen vorkommen. Bei Wikipedia-basierten Verfahren werden so auch wahrscheinlich vom Benutzer nicht intendierte Titel wie „City of“ (Episodenname einer Fernsehserie) oder „In New York“ (Album eines Jazzmusikers) als Konzept erkannt.

Wikipediatitel im Webis Query Segmentation Corpus

Im Webis Query Segmentation Corpus gibt es insgesamt 36 886 Anfragen, die Wikipediatitel mit mindestens zwei Wörtern enthalten. Dabei wurden Titel ab einer Länge von zwei Wörtern berücksichtigt. In 17 658 Anfragen kommen dabei mindestens zwei Wikipediatitel vor. Die Anfragen mit überlappenden Wikipediatiteln können in Anfragen mit sich gegenseitig überlappenden Titeln und solche, bei denen ein Wikipediatitel den anderen einschließt, unterschieden werden. Daneben gibt es 7432 Fälle, in denen sich mindestens zwei Wikipediatitel, die weder verschachtelt noch überlappend sind, in einer Anfrage befinden.

Unterschiedliche Wikipediatitel als Hinweis auf Mehrdeutigkeit

Beim Betrachten der Verteilung der Anfragen mit überlappenden Wikipediatiteln fällt auf, dass diese im Vergleich zur Verteilung im gesamten Korpus, einen höheren Anteil unsicherer Anfragen enthält. Die Vermutung liegt nahe, dass Anfragen, bei denen sich Menschen beim Segmentieren uneinig sind, häufig mehr als nur eine Interpretationsmöglichkeit haben, was zu einer unsicheren Stimmverteilung führt. Die Anfragen mit überlappenden Wikipediatiteln können in Anfragen mit sich gegenseitig überlappenden Titeln und solche, bei denen ein Wikipediatitel den anderen einschließt, unterschieden werden.

Im Webis Query Segmentation Corpus gibt es 3642 Anfragen, bei denen sich Titel aus der Wikipedia so überlappenden, dass der eine Titel mindestens ein Wort enthält, das in dem anderen Titel nicht vorkommt und umgekehrt:

$$\overbrace{\texttt{you are}\underbrace{\texttt{love}}}^{\text{WT1}}\underbrace{\texttt{song}}_{\text{WT2}}.$$

Außerdem gibt es 6584 Suchanfragen, bei denen die Wortefolge des einen Wikipediatitels komplett in einem längeren Titel vorhanden ist:

$$\overbrace{\underbrace{\texttt{iraq war}}_{\text{WT2}}\texttt{casulties}}^{\text{WT1}}.$$

Für Suchanfragen, die sich gegenseitig überlappende Wikipediatitel enthalten, kann das Vorhandensein unterschiedlicher, sich überschneidender Konzepte angenommen werden. Die als unsicher eingestuften Anfragen mit sich gegenseitig überlappenden Wikipediatiteln ab einer Titellänge von zwei Wörtern, die 38,41 % ausmachen (vgl. Abb. 3.1), sind bei der Suche nach mehrdeutigen Anfragen besonders interessant.

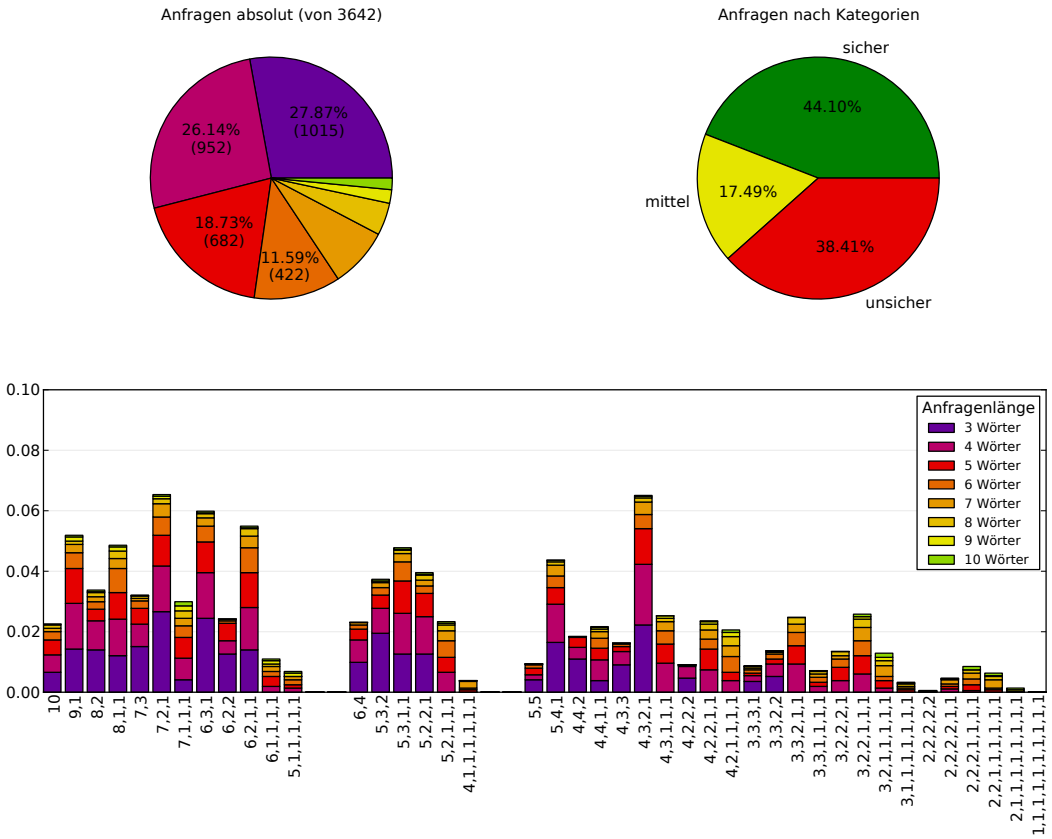

Abbildung 3.1: Die Verteilung der Anfragen mit überlappenden Wikipediatiteln

Filtern von Wikipediatiteln

Da Wikipedia teilweise auch Artikel enthält, deren Inhalt nur einer relativ kleinen Benutzerschaft bekannt ist, ergibt sich das Problem, dass auf der Suche nach mehrdeutigen Anfragen mit Hilfe von überlappenden Wikipediatiteln viele Titel „erkannt“ werden, die dazugehörige Bedeutung aber so nicht vom Anfragensteller gemeint war. Außerdem erschweren bei Nicht-SNP-Anfragen Wikipediaeinträge zu Konzepten wie „is a“ (Vererbungsprinzip beim objektorientierten Programmieren) und „how to“ (Anleitung zur Lösung eines bestimmen Problems) die Suche. So ist es zum Beispiel sehr unwahrscheinlich, dass in der Anfrage `how to do telekinesis` mit `to do` das Konzept Zeitmanagement gemeint ist, für das ein Artikel in der Wikipedia mit diesem Titel existiert. Solche für die Suche nach mehrdeutigen Anfragen nicht geeigneten Titel müssen deshalb herausgefiltert werden. Dabei werden diejenigen Titel

Wikipediatitel	Stimmen	Anfragen	Stimmen pro Anfrage
how to	4628	544	8.51
for sale	2394	461	5.19
city of	702	117	6.00
how do	543	60	9.05
is a	536	59	9.08
for kids	432	51	8.47
to do	427	56	7.62
real estate	418	246	1.70
for rent	374	77	4.86
in california	340	37	9.19
do i	307	37	8.30
do you	298	35	8.51
homes for sale	290	70	4.14
how do i	264	29	9.10
to be	251	34	7.38
can i	248	27	9.19
how many	243	27	9.00
the best	218	26	8.38

Tabelle 3.1: Häufig nicht erkannte Wikipediatitel, sortiert nach der Anzahl der Menschen, die den jeweiligen Titel nicht segmentiert haben.

aussortiert, die im Durchschnitt sieben mal oder öfter nicht segmentiert worden sind.

Es wurden nach diesem Kriterium 2163 Wikipediatitel herausgefiltert, sodass noch 23 317 übrig bleiben. Demnach sind es noch 2763 Anfragen mit überlappenden Wikipediatiteln, davon 42,67 % als „unsicher“ eingestuft. Anschließend wurden 1179 „unsichere“, 497 „mittlere“ und 1087 „sichere“ Anfragen manuell auf Mehrdeutigkeit überprüft.
Von ursprünglich 6584 Anfragen, bei denen ein Wikipediatitel einen anderen einschließt, bleiben noch 4778 Suchanfragen übrig, wovon 38,66 % als „unsicher“ einzuordnen sind.

Mehrdeutige Anfragen mit überlappenden Wikipediatiteln

Nachdem die 2763 Anfragen mit sich gegenseitig überlappenden Wikipediatiteln einzeln überprüft wurden, bleiben nur 25 mehrdeutige Anfragen übrig. Davon 17 Anfragen, die als „unsicher“ eingestuft sind, sechs als „mittel“ und drei als „sicher“. Offensichtlich ist die Unsicherheit beim Segmentieren nicht

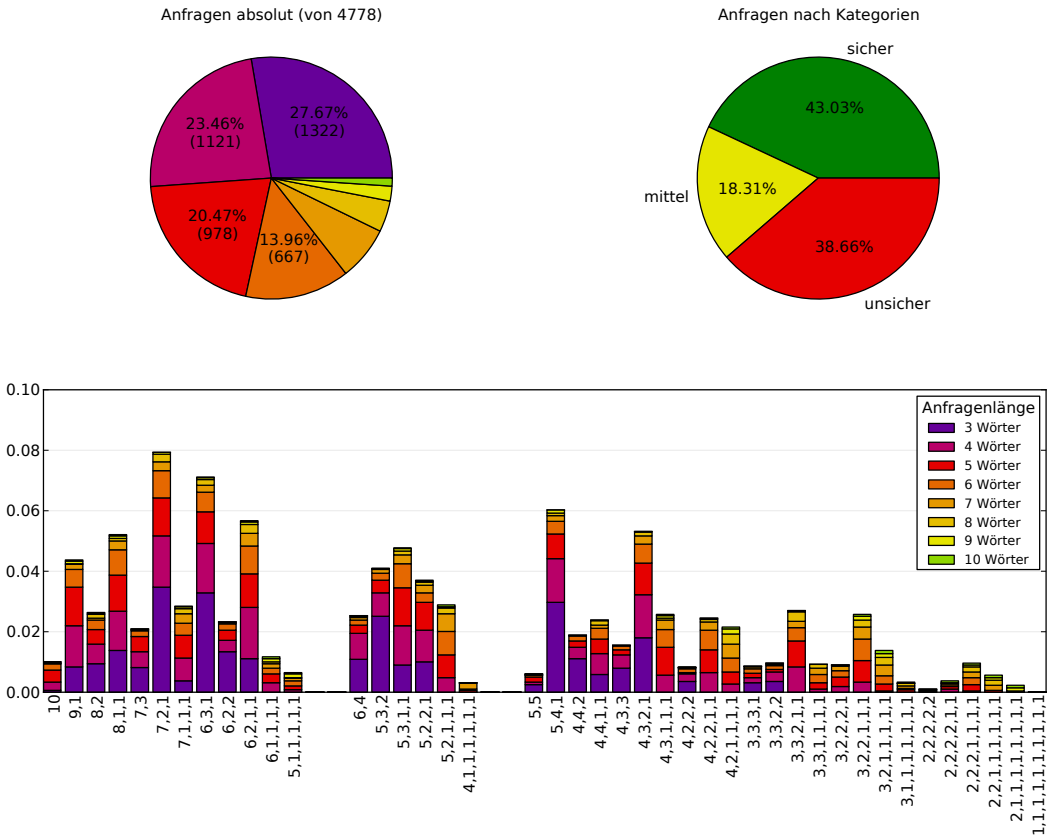

Abbildung 3.2: Die Verteilung der Anfragen, bei denen ein Wikipediatitel von einem anderen eingeschlossen wird.

auf die überlappenden Wikipediatitel zurückzuführen, da nur 1,44 % der Anfragen nach manueller Überprüfung wirklich als mehrdeutig bezeichnet werden können. Bei den „unsicheren“ Anfragen mit Wikipediatiteln ineinander konnten beim einzelnen Untersuchen keine mehrdeutigen Suchanfragen gefunden werden.

Drei- und Vier-Wort-Anfragen

Drei- und Vier-Wort-Anfragen bestehen relativ häufig aus SNPs. Da Suchmaschinen schon gut mit SNPs umgehen können, macht es keinen Unterschied für die Retrieval Performance, ob sie segmentiert werden oder nicht. Deshalb sollten SNPs lieber nicht segmentiert werden, um die Qualität der Suchergebnisse nicht zu verschlechtern [HPBS12]. Außerdem lassen sich im Webis Query

Segmentation Corpus für die Wortanzahlen drei und vier mehr Anfragen in die Kategorie „sicher“ einordnen als in die anderen beiden zusammen.

Unsicherheit und Existenz mehrerer Wikipediatitel

Es bleibt noch zu klären, ob die Unsicherheit der segmentierenden Menschen daran liegt, dass eine Anfrage mehrere unterschiedliche Wikipediatitel enthält. Dazu wird betrachtet, wie die enthaltenen Wikipediatitel von den Arbeitern segmentiert wurden. Dabei wird unterschieden, ob überhaupt eine nennenswerte Anzahl von Titeln segmentiert wurde und wie dann das Zahlenverhältnis zwischen den verschiedenen segmentierten Wikipediatiteln ist. Dabei wird folgende Einteilung vorgenommen:

Kategorie	Verteilung
sicher	’10,0’, ’9,0’, ’9,1”, ’8,0’, ’8,1’, ’8,2’, ’7,0’, ’7,1’, ’7,2’, ’7,3’, ’6,0’, ’6,1’, ’6,2’
mittel	’6,3’, ’5,2’
unsicher	’6,4’, ’5,5’, ’5,4’, ’5,3’, ’4,4’, ’4,3’
verbleibend	’5,0’, ’5,1’, ’4,0’, ’4,1’, ’3,0’, ’4,2’, ’3,3’, ’3,2’, ’3,1’, ’2,2’, ’2,1’, ’2,0’, ’1,1’, ’1,0’, ’0,0’

Segmentierung überlappender Wikipediatitel durch Annotatoren

Für einen sinnvollen Vergleich, welcher Wikipediatitel wie häufig segmentiert wurde, werden nur Anfragen in die Kategorien „sicher“, „mittel“ und „unsicher“ eingeteilt, bei denen beide Wikipediatitel zusammen mindestens sieben mal segmentiert wurden (oder ein Titel sechs mal und der andere gar nicht).

Es zeigt sich bei Anfragen, in denen ein Wikipediatitel den anderen einschließt, dass die Unsicherheit der Arbeiter beim Segmentieren nicht der Tatsache geschuldet ist, dass die Anfrage mehrere sich überlappende Wikipediatitel enthält. Nur 25,61 % (677 mal) der Anfragen können der Kategorie „unsicher“

zugeordnet werden (vgl. Tab. A.4). Noch deutlicher zeigt sich dies, wenn man die Anfragen mit den Wortlängen drei und vier nicht mit betrachtet. Hier sind dann mit einer Anzahl von 293 die Anfragen, bei denen die Segmentierung der enthalteten Wikipediatitel als „sicher“ einzustufen sind, fast gleich denen mit „unsicher“ segmentierten Wikipediatiteln, die 294 betragen. Der größte Teil entfällt jedoch mit 58,25 % (1540 Anfragen) auf solche Fälle, die aufgrund geringer Segmentierung der Wikipediatitel keiner Kategorie zugeordnet werden können. (vgl. Abb. 3.3).

Ähnlich verhält es sich mit Anfragen, bei denen sich Wikipediatitel gegenseitig überlappen. Die Anfragen, die keiner Kategorie zugeordnet werden können, machen sogar 70,70 % (1151 Anfragen) aus (vgl. Tab. A.5). Beachtet man dabei die 3- und 4-Wort-Anfragen nicht, überwiegen sogar die Anfragen, die aufgrund der segmentierten Wikipediatitel als „sicher“ eingestuft wurden mit 163 gegenüber den „unsicheren“, die nur eine Anzahl von 112 haben (vgl. Abb. 3.3).

Dass „sichere“ Anfragen mit überlappenden Wikipediatiteln auch eine „sichere“ Segmentierung der Wikipediatitel aufweisen, ist aufgrund einer relativ kleinen Anzahl von unterschiedlichen Segmentierungen bei „sicheren“ Anfragen nicht überraschend. So sind es bei Anfragen mit Wikipediatiteln ineinander 75 % aller Fälle (vgl. Tab. A.7), bei denen Wikipediatitel sicher segmentiert werden und bei Anfragen, in denen sich Wikipediatitel gegenseitig überlappenden, immerhin noch 61,9 % (vgl. Tab. A.6).

Segmentierung sich einschließender Wikipediatitel

Anschließend wurde analysiert, welcher Wikipediatitel bei Anfragen mit geschachtelten Titeln gewählt wurde. Es wurden nur Anfragen mit einer Mindestlänge von fünf Wörtern berücksichtigt. Bei solchen Anfragen gibt es immer einen kurzen Titel und einen längeren, der den kürzeren einschließt. Es zeigt sich deutlich, dass, egal in welcher Kategorie, die Segmentierer den längeren Wikipediatitel häufiger segmentieren als den kürzeren. Insgesamt wird in 66,71 % aller Fälle der längere Wikipediatitel am häufigsten gewählt. Das bestätigt die Entscheidung von Hagen et al., deren Segmentierungsalgorithmus in solchen Fällen durch die Verwendung von N-Grammen multipliziert mit einem Wichtungsfaktor längere Wikipediatitel bevorzugt [HPBS12]. Der kürzere Titel erscheint in 21,58 % aller betreffenden Anfragen als das Segment, das am meisten segmentiert wird und mit 11,71 % werden beide Wikipediatitel gleich oft als Segment gruppiert. Durch die gewählte Einteilung der Kategorien bei der Segmentierung der Wikipediatitel erscheinen in den Kategorien „sicher“ und „mittel“ keine Anfragen, bei denen beide Titel gleich oft segmentiert werden. Solche Anfragen gibt es vor allem in den Kategorien „unsicher“

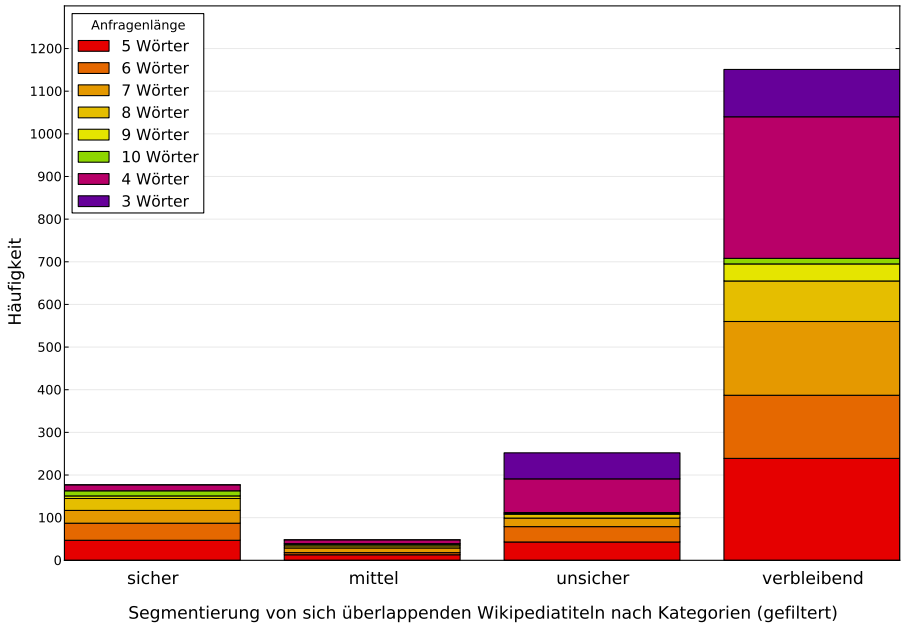

Abbildung 3.3: Die Segmentierung von sich gegenseitig überlappenden Wikipediatiteln in unsicheren Anfragen getrennt nach Kategorien. In 70,7 % (1151 Fälle) der Anfragen sind beide Wikipediatitel zusammen nicht öfter als sechs Mal segmentiert worden.

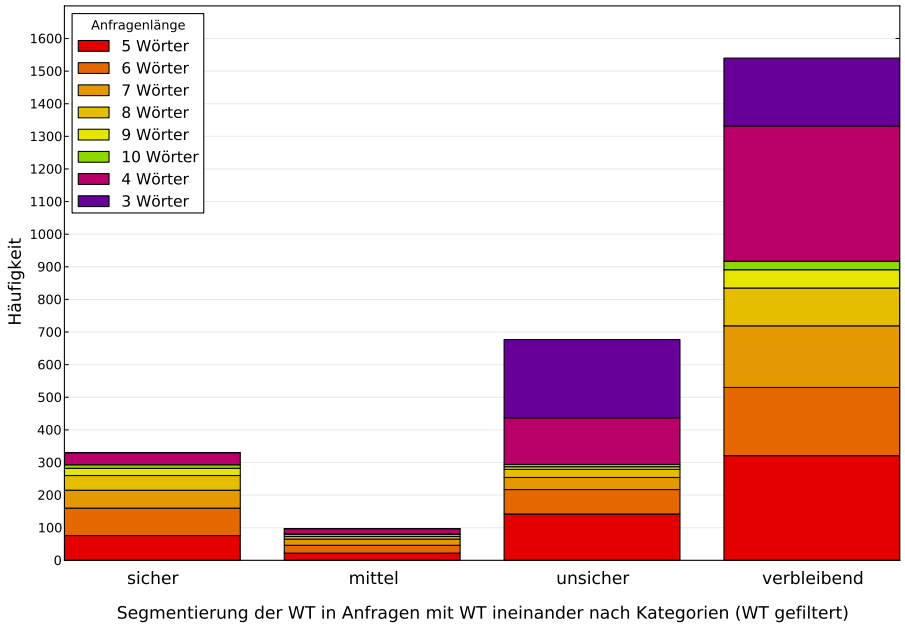

Abbildung 3.4: Die Segmentierung in unsicheren Anfragen, in denen ein Wikipediatitel einen anderen einschließt, getrennt nach Kategorien. In 58,25 % (1540 Fälle) der Anfragen sind beide Wikipediatitel zusammen nicht öfter als sechs Mal segmentiert worden.

und „verbleibend“, in denen die Verteilungen „5,5“, „4,4“ bzw. „3,3“, „2,2“, „1,1“ und „0,0“ zu finden sind. Es ist deutlich erkennbar, dass in den Anfragen, in denen ein Wikipediatitel sicher segmentiert wurde, mit 82,35 % am häufigsten der längere Wikititel gewählt wurde. (vgl. Tab. 3.2).

Segmentierung sich gegenseitig überlappender Wikipediatitel

Außerdem wurde die Segmentierung bei Anfragen mit fünf Wörtern Mindestlänge, in denen sich Wikipediatitel gegenseitig überlappen, betrachtet. Hierbei ist besonders von Interesse, ob die Segmentierer mehrheitlich den Wikipediatitel mit höherer oder niedrigerer N-Gramm-Häufigkeit wählten. Es zeigt sich, dass insgesamt in 64,45 % der betreffenden Anfragen sich die Mehrheit für den Wikititel mit der höheren N-Gramm-Häufigkeit entschied. In nur 19,68 % der Fälle favorisierte die Mehrzahl den Titel mit der niedrigeren N-Gramm-Häufigkeit, 15,87 % der Anfragen wiesen einen Gleichstand für die Entscheidung für einen der beiden Wikipediatitel auf. Am deutlichsten ist das bei den Anfragen zu sehen, bei denen von den Segmentierern klar ein Wikipediatitel bevorzugt wurde und die deshalb in der Kategorie „sicher“ zu finden sind. Hier werden in 84,82 % aller entsprechenden Anfragen die Wikipediatitel mit der höchsten N-Gramm-Häufigkeit am deutlichsten mehrheitlich segmentiert (vgl. Tab. 3.3).

Verschiedene Suchergebnisse bei unterschiedlicher Segmentierung

Um mehrdeutige Anfragen durch unterschiedliche Suchergebnisse für verschiedene Segmentierungen derselben Anfrage zu finden, wurden Ergebnislisten der Suchmaschine Bing verwendet. Dafür wurden die im Query Segmentation Corpus vorhandenen Segmentierungen als Anfrage an die Bing-API gestellt. Unterschiede in den Ergebnissen für verschiedene Segmentierungen einer Anfrage könnten ein Hinweis auf Mehrdeutigkeit sein, die durch Segmentieren beseitigt wird. Es wurden die ersten zehn Ergebnisse der Suchergebnisliste analyisert. Jedoch konnte unter den Anfragen, bei denen die Ergebnisse sich für verschiedene Segmentierungen unterschieden, keine mehrdeutigen Anfragen gefunden werden.
Der Algorithmus von Bing ist, einer „Black Box“ gleich, nicht bekannt. Zwar scheinen sich teilweise die Suchergebnisse für unterschiedliche Segmentierungen einer Anfrage zumindest in der Reihenfolge zu unterscheiden. Jedoch ist nicht klar, wie groß der Einfluss der Aufteilung der einzelnen Segmente auf die präsentierten Ergebnisse ist.

Segm.\WT	sicher			mittel			unsicher			verbleibend		
	l	k	g	l	k	g	l	k	g	l	k	g
sicher	800	154	0	18	9	0	5	4	0	85	44	161
mittel	167	30	0	29	10	0	58	12	3	124	42	32
unsicher	218	70	0	59	28	0	118	99	65	489	200	123
Summe	1185	254	0	106	47	0	181	115	68	698	286	316

Tabelle 3.2: Die Tabelle gibt an, wie oft in einer Anfrage mit Wikipediatiteln ineinander der kürzere (k) oder längere (l) häufiger Wikipediatitel gewählt wurde. Außerdem kam es auch vor, dass beide Wikipediatitel gleich oft gewählt wurden (g). Dabei wird zwischen Anfragen, die ihren Segmentierungen nach „sicher“, „mittel“ und „unsicher“ sind (Zeilen) und nach Segmentierung der Wikipediatitel (Spalten) unterschieden.

Segm.\WT	sicher			mittel			unsicher			verbleibend		
	h	n	g	h	n	g	h	n	g	h	h	g
sicher	354	53	0	6	2	0	1	0	0	49	26	111
mittel	59	7	0	5	2	0	16	4	0	50	19	19
unsicher	73	27	0	18	9	0	32	26	13	267	109	86
Summe	486	87	0	29	13	0	49	30	13	366	154	216

Tabelle 3.3: Die Zeilen sind aufgeteilt in die Kategorien nach Segmentierungen, die Spalten sind nach „Sicherheit“ der Segmentierung der Wikipediatitel geordnet. Hier wird jeweils unterschieden zwischen dem Fall, dass die Segmentierer bei überlappenden Wikipediatiteln sich mehrheitlich für die Segmentierung mit der höchsten N-Gramm-Häufigkeit (h) oder die niedrigere (n) entscheiden. Außerdem kam es vor, dass sich jeweils gleich viele Menschen für einen Wikipediatitel entscheiden (g).

3.2 Analyse der mehrdeutigen Suchanfragen

Im Folgenden werden die gefundenen mehrdeutigen Anfragen unter Verwendung von Wikipediatiteln genauer untersucht.

Anzahl mehrdeutiger Anfragen

Mit gerade einmal 0,94 % (25 von 2763) der Anfragen mit überlappenden Wikipediatiteln, nachdem nicht relevante Wikipediatitel entfernt worden sind, liegt die Anzahl mehrdeutiger Anfragen viel niedriger als erwartet. Bei einem Anteil von 29,14 % (15 570) unsicherer Anfragen im Korpus war die Annahme, Mehrdeutigkeiten seien eine Hauptursache für die Uneinigkeit. Der Anteil der gefundenen Suchanfragen macht im Webis Query Segmentation Corpus insgesamt nur 0,049 % (25 von insgesamt 53 432) aus.

Qualität der gefundenen mehrdeutigen Anfragen

Neben dem Problem, dass das manuelle Filtern der mehrdeutigen Anfragen durch subjektive menschliche Sichtweisen beeinflusst ist, sind nicht alle ermittelten Anfragen von gleicher Qualität. Beispielsweise ist der inhaltliche Unterschied zwischen `louisiana state|animal` und `louisiana|state animal` viel größer als `you are love|song` und `you are|love song`. Bei dem einen besteht der Unterschied höchstwahrscheinlich zwischen der Suche nach dem Logo des US-Bundesstaates Louisiana und dem Interesse für in diesem Bundesstaat lebende Tiere. Bei dem anderen unterscheidet sich bei der einen Segmentierung höchstens der Name und die Einstufung des Werkes als Liebeslied von der anderen, in beiden wird aber nach einem Musikstück gesucht.

Manche Anfragen könnten auch durch leichte Veränderungen noch an Mehrdeutigkeit dazugewinnen. Die Suchanfrage `skyline drive in` könnte zum Beispiel mit dem zusätzlichen Wort `virginia` noch deutlicher eine Segmentierung nötig machen, um zwischen einem Autokino und einer Straße durch einen Nationalpark im selben US-Bundesstaat zu unterscheiden.

Anfragenlänge und SNPs

Bis auf eine Ausnahme mit fünf Wörtern bestehen die gefundenen Anfragen vor allem aus drei sowie aus vier Wörtern. Ähnlich der Verteilung im Korpus, in welchem kurze Anfragen vor allem „Strict Noun Phrases“ (SNP) sind, finden sich unter den mehrdeutigen 22 SNP-Anfragen und nur drei, die keine SNP sind. Eine Begründung dafür könnte sein, dass Anfragen mit Wortlänge fünf und mehr nur sehr selten im Korpus vorkommen. Außerdem könnte es sein,

dass generell bei Anfragen mit mehr Wörtern dadurch, dass diese mehr Sätzen natürlicher Sprache entsprechen, weniger Mehrdeutigkeiten vorhanden sind.

Verteilung der Segmentierungen

Die meisten Anfragen fallen aufgrund der Verteilung der Segmentierungen in die Kategorie „unsicher“. Hier liegt die Vermutung nahe, dass die Mehrdeutigkeit mit verantwortlich für die Wahl unterschiedlicher Segmentierungen verschiedener Segmentierer bei der selben Anfrage ist. Alle vorhandenen mehrdeutigen Anfragen wurden durch das Überprüfen von Suchanfragen mit sich gegenseitig überlappenden Wikipediatiteln gefunden.

Ursachen für Unsicherheit bei mehrdeutigen Anfragen

Bei der Analyse, wie oft die Wikipediatitel jeweils segmentiert wurden, zeigte sich, dass nur in 11 von 25 Fällen zwei Titel zusammen mindestens sieben mal segmentiert wurden. In 7 der 11 Anfragen ist die Verteilung als „unsicher“ einzustufen, hier liegt die Unsicherheit an den sich überschneidenden inhaltlichen Konzepten. Obwohl die mehrdeutigen Anfragen durch sich ausschließlich überlappende Wikipediatitel gefunden wurden, scheint die Ursache für die Unsicherheit der Segmentierer bei mindestens drei Anfragen daran zu liegen, dass es auch einen Wikipediatitel gibt, der einen anderen einschließt. Das ist beispielsweise bei der Anfrage `washington mutual bank` der Fall. Hier schließt der Titel `washington mutual bank` die beiden sich überlappenden anderen Titel `mutual bank` und `washington mutual` ein.
Bei den verbleibenden restlichen Anfragen bleibt die Ursache für die Uneinigkeit der Segmentierer unklar, es kann jedoch nicht ausgeschlossen werden, dass die von Wikipediatiteln repräsentierten Konzepte zumindest teilweise ein Grund sein können.

Anfrage	WT1	WT2	WT3
studio art therapy	art therapy (3)	studio art (2)	-
african american album cover art	album cover (3)	cover art (1)	album cover art (3)
internet business opportunities	internet business (2)	business opportunities (4)	-
homeland security systems	security systems (4)	homeland security (3)	-
reddy ice plant	reddy ice (2)	ice plant (0)	-
natural hair products	hair products (5)	natural hair (3)	-
georgia state parks	state parks (3)	georgia state (1)	-
the n game	the n (0)	n game (2)	-
manhattan beach music	beach music (3)	manhattan beach (6)	-
free jazz guitar licks	jazz guitar (5)	free jazz (1)	-
sony music software	sony music (5)	music software (2)	-
online money management class	money management (4)	online money (0)	-
manchester union leader	union leader (5)	manchester union (3)	manchester union leader (1)
automobile blue book value	book value (1)	blue book (4)	blue book value (5)
skyline drive in	drive in (4)	skyline drive (1)	-
arizona state parks	state parks (7)	arizona state (1)	arizona state parks (1)
international law degree	law degree (4)	international law (2)	-
striped bass river basin	striped bass (5)	bass river (2)	river basin (2)
you are love song	you are love (3)	love song (6)	-
free cell games	cell games (4)	free cell (1)	-
washington mutual bank online	mutual bank (3)	washington mutual (2)	washington mutual bank (4)
best small business college	business college (1)	small business (4)	-
single family assistance	single family (2)	family assistance (1)	-
louisiana state animal	louisiana state (5)	state animal (4)	-
bible study guide	bible study (4)	study guide (3)	-

Tabelle 3.4: Die gefundenen mehrdeutigen Anfragen mit enthaltenen Wikipediatiteln (WT1, WT3 und WT3). In Klammern ist angegeben, wie oft der Titel segmentiert wurde. In der Anfrage `african american album cover art` gibt es mit `african american` noch einen vierten Wikipediatitel, der von neun Segmentierern erkannt wurde und bei dem deshalb davon ausgegangen werden kann, dass er keine Auswirkung auf die Mehrdeutigkeit hat.

Kapitel 4

Säuberung des Korpus

Der im Folgenden beschriebene Ansatz, um ein Korpus zu reinigen, besteht darin, äußerlich verschiedene, inhaltlich aber gleiche Segmentierungen zu erkennen und zu vereinigen. Um ähnliche Segmente zu erkennen, werden die Stellen, an denen ein Segment endet und ein neues beginnt, verglichen. Anschließend wird die Veränderung der Segmentierungsverteilung für 5-Wort-Anfragen betrachtet.

4.1 Unterschiedlichkeit von Segmentierungen

Bei 71,18 % (38 035 absolut) der Anfragen im Query Segmentation Corpus kommt es vor, dass von den menschlichen Segmentierern mindestens einmal keine Segmentierung vorgenommen wurde, d.h. jedes Wort ein einzelnes Segment darstellt. Jedoch ist nur in 10 461 Fällen die von den meisten Menschen bevorzugte Segmentierung die Nicht-Segmentierung. Oft unterscheiden sich so inhaltlich gleiche, aber von der Gruppierung der einzelnen Wörter her unterschiedliche Segmentierungen, nur durch wenige Trennstellen. Deshalb können solche Segmentierungen durch das Weglassen von Trennstellen vereinigt werden.

$$breaks(S_1) \cup breaks(S_2) = breaks(S_2)$$
$$breaks(S_1) - breaks(S_2) = \emptyset$$

So können beispielsweise bei der Anfrage `graffiti fonts alphabet` die ersten beiden Segmentierungen durch das Weglassen einer Trennstelle vereinigt werden. So lassen sich die ursprünglich drei unterschiedlichen Segmentierungen:

	Segmentierung	Stimmen
S_1	`graffiti fonts|alphabet`	5
S_2	`graffiti|fonts|alphabet`	3
S_3	`graffiti fonts alphabet`	2

zu nur noch zwei verschiedenen Segmentierungen zusammenfassen:

	Segmentierung	Stimmen
S_1	`graffiti fonts|alphabet`	8
S_2	`graffiti fonts alphabet`	2

.

Dahingegen wird angenommen, dass Segmentierungen unterschiedlich gemeint sind, wenn Trennstellen sich so unterscheiden, dass eine Segmentierung Trennstellen besitzt, die die andere nicht hat, und umgekehrt.

$$breaks(S1) \cap breaks(S2) \neq breaks(S1)$$
$$breaks(S1) - breaks(S2) \neq \emptyset$$

Demnach ist bei den Segmentierungen der Anfrage `maui beach rentals` keine Vereinigung möglich:

	Segmentierung	Stimmen
S_1	`maui beach|rentals`	7
S_2	`maui|beach rentals`	3

.

Die leere Segmentierung, bei der jedes Wort ein einzelnes Segment bildet, ist vom Vereinigen ausgenommen. Diese Form der Segmentierung unterscheidet sich grundlegend von den anderen, weil sich der Segmentierer dazu entschieden hat, gar nicht zu segmentieren. Deshalb soll die leere Segmentierung nicht durch das Vereinigen mit einer anderen Segmentierung verschwinden.

4.2 Vereinigen ähnlicher Segmentierungen

Eine Segmentierung kann mit einer anderen, häufiger gewählten Segmentierung vereinigt werden, wenn sich die häufiger gewählte durch Weglassen einer bestimmten Anzahl Trennstellen bei der seltener gewählten erzeugen lässt.

Ob zwei Segmentierungen als gleich betrachtet werden können, hängt dabei von der Anfragenlänge ab. Je länger die Anfrage ist, desto mehr Trennstellen

Algorithm 1 Vereinigung durch Weglassen von Trennstellen

```
▷ D: Dict mit Segmentierungslisten
▷ S: Liste mit Segmentierungen
for all S in D do
    function VEREINIGUNG(D)
        for i ← n − 1, 0 do
            for j ← n, i + 1 do
                if S(i) und S(j) gleich then
                    vereinige S(i) und S(j)
                    sortiere S nach Stimmenanzahl
                    return VEREINIGUNG(D)
                end if
            end for
        end for
    end function
end for
```

können weggelassen werden, um die gleiche Segmentierung zu erhalten. Bei 3- und 4-Wort-Anfragen ist das Entfernen einer Trennstelle erlaubt, bei Länge fünf und sechs dürfen bis zu zwei solcher Stellen weggelassen werden, bis hin zu vier entfernbaren Trennstellen bei Anfragen mit neun und zehn Wörtern. Außerdem kann bei zwei ähnlichen Segmentierungen nur zugunsten der häufiger gewählten vereinigt werden, weil sonst zu viele Stimmen unterschlagen werden. Jedoch besteht bei dieser Methode die Gefahr, dass zwei sich nicht ähnelnde Segmentierungen jeweils mit einer Segmentierung, die mehr Stimmen erhalten hat, vereinigt werden. Bei einer mehrdeutigen Anfrage würden so Interpretationsmöglichkeiten verloren gehen. Beispielsweise verschwinden bei der Anfrage `unisource energy services` so die Segmentierungen 2 und 3, obwohl sie sich nicht ähnlich sind:

	Segmentierung	Stimmen
S_1	`unisource energy services`	4
S_2	`unisource energy|services`	3
S_3	`unisource|energy services`	2
S_4	`unisource|energy|services`	1

Um das zu verhindern und die abgebildeten Meinungen, wie in dem jeweiligen Fall richtig zu segmentieren ist, nicht zu verfälschen, dürfen sich gegenseitig unähnliche Segmentierungen nicht zusammen in einer anderen Segmentierung aufgehen. Nachdem das Vereinigen unter Beachtung dieser Bedingung durchgeführt wurde, weist die „gesäuberte“ Variante des Webis Query Segmentation

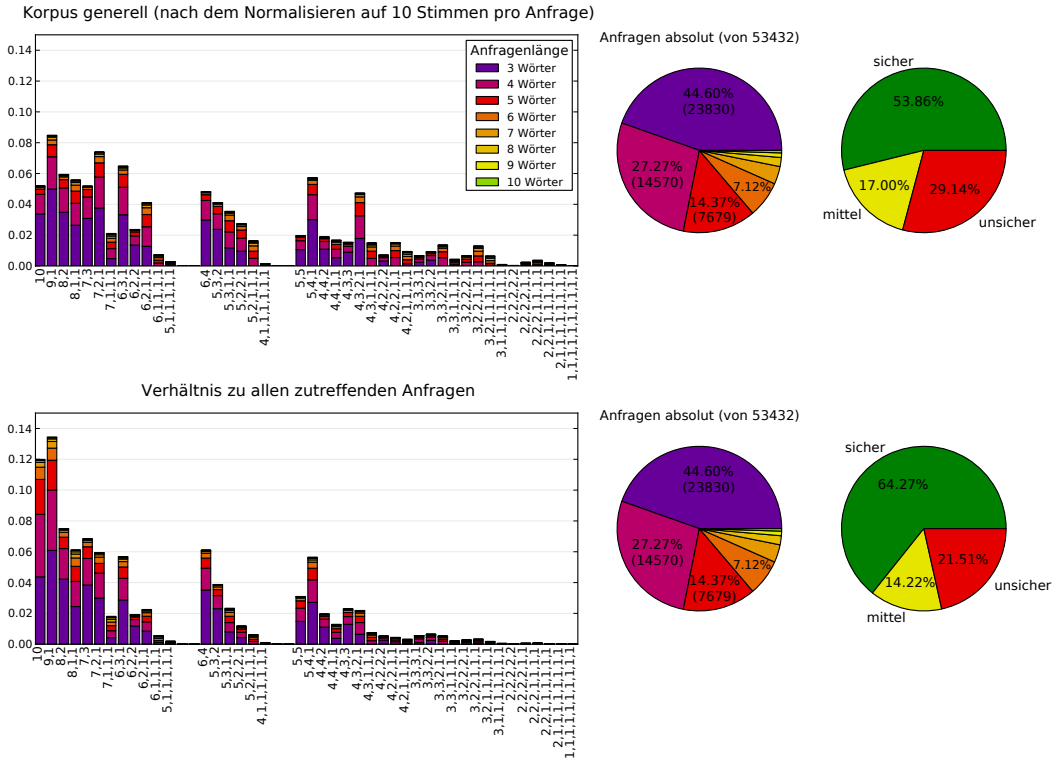

Abbildung 4.1: Gegenüberstellung der Anfragenverteilung im Webis Query Segmentation Corpus generell (oben) und der Verteilung der Anfragen nachdem ähnliche Segmentierungen vereinigt wurden (unten). Der Anteil der sicheren Anfragen konnte so erhöht und jener der als unsicher eingestuften Anfragen verringert werden.

Corpus mit 64,27 % eine um 10,41 % verbesserte Rate sicherer Anfragen auf. Der Anteil unsicherer Anfragen konnte um 7,63 % auf 21,51 % verringert werden (vgl. Abb. 4.1).

Veränderung der Segmentierungsverteilung

Beim Vereinigen von zwei unterschiedlichen Segmentierungen wird über das Entfernen von Trennstellen, also durch das Verbinden zweier Segmente, ein neues gebildetet, bis die Segmente der einen Segmentierung genau denen der anderen entsprechen.
Im Folgenden wird die Veränderung der Segmentierungsverteilung bei Anfragen, die aus fünf Wörtern bestehen, betrachtet. Anfragen mit den Längen drei und vier sind zu kurz und nicht so interessant, weil bei ihnen eine Segmentie-

rung hinsichtlich der Retrieval Performance keine Vorteile bringt. Es gibt im Webis Query Segmentation Corpus 7679 5-Wort-Anfragen, das sind mehr als die restlichen Anfragen mit sechs bis zehn Wörtern zusammen.

Am häufigsten tritt die leere Segmentierung auf, bei der alle Wörter der Suchanfrage als einzelnes Segment vorhanden sind. In 19,78 % der Fälle (12 683 von 76 790), in denen ein Segmentierer die Anfragewörter gruppieren musste, wurde die Anfrage also unsegmentiert abgeschickt, sodass die Segmentierung aus fünf 1-Wort-Segmenten besteht. Weil diese Segmentierung sich also grundsätzlich von den anderen dadurch unterscheidet, dass nicht segmentiert wurde, werden solche Fälle beim Vereinigen von ähnlichen Segmentierungen nicht mit berücksichtigt.

Segmentanzahl

Es werden Segmentierungen mit drei Segmenten (0110, 0011, 1100, 0101, 1010, 1001) bevorzugt. Auf sie entfallen 32,02 % aller Stimmen. Beim Vereinigen steigt dieser Anteil leicht auf 32,86 %. Segmentierungen, die zwei Segmente besitzen (1000, 0100, 0010, 0001) können durch das Unifizieren am meisten zulegen, sie machen vor dem Vereinigen 19,75 % aller 5-Wort-Segmentierungen aus, danach 27,36 %.

Die Segmentierung, die mit 11,84 % die meisten Stimmen (9089 von 76 790 aller 5-Wort-Anfragen) im Korpus vor dem Vereinigen hat, besitzt am Anfang und am Ende je ein 2-Wort-Segment: $w_1w_2|w_3|w_4w_5$ (0110). Durch das Vereinigen kommen noch einmal 2500 Stimmen dazu, während 1048 (11,53 %) zu anderen Segmentierungen wechseln.

Wechsel von Segmentierungen

Die Segmentierung der Form $w_1w_2w_3|w_4w_5$ (0010) gewinnt durch das Vereinigen am meisten Stimmen hinzu und kann seinen Anteil um 5,1 % steigern. Vor allem wechseln die Segmentierungsformen $w_1|w_2w_3|w_4w_5$ (1010) und $w_1w_2w_3|w_4|w_5$ (0011) mit drei Segmenten zu dieser Art der Segmentierung (vgl. Abb. 4.2). Auch die Segmentierung mit dem Schema 0100 kann 3,16 % dazugewinnen. Ebenfalls zulegen kann die beliebteste 0110-Segmentierung. Hier beträgt die Zunahme jedoch nur 2,26 %, weil einige Stimmen zur 0010-Segmentierung wechseln.

An erster Stelle veringert sich der Anteil von Segmentierungen mit vier Segmenten. Mit 4,06 % verliert die 0111-Segmentierung die meisten Stimmen, die 1101-Segmentierung 1,26 %. Das könnte ein Hinweis darauf sein, dass viele solcher Segmentierungen von den Menschen ungünstig gewählt wurden.

Segmentierung	Stimmen ursprünglich		Stimmen vereinigt	
	absolut	in Prozent	absolut	in Prozent
1111	12683	19.78 %	12683	19.78 %
0110	9089	14.18 %	10541	16.44 %
1110	9069	14.15 %	6211	9.69 %
0111	7380	11.51 %	4776	7.45 %
0010	6206	9.68 %	9463	14.76 %
0100	5089	7.94 %	7119	11.10 %
0011	5057	7.89 %	4568	7.13 %
1100	4544	7.09 %	4499	7.02 %
1011	3153	4.92 %	2322	3.62 %
0101	2752	4.29 %	2689	4.19 %
1101	2573	4.01 %	1761	2.75 %
0001	2481	3.87 %	2920	4.55 %
1010	2200	3.43 %	2042	3.19 %
0000	2178	3.40 %	2792	4.36 %
1000	1392	2.17 %	1511	2.36 %
1001	944	1.47 %	893	1.39 %

Tabelle 4.1: Diese Verteilung der 5-Wort-Anfragen ergibt sich vor bzw. nach dem Unifizieren. Der Anteil der leeren Segmentierung 1111 ist dabei unverändert geblieben, weil sie beim Vereinigen nicht mit einbezogen wird.

Verteilung		Wechselsegmentierung											
		0010	0100	0000	0110	1100	1010	1000	0011	0101	0001	1001	verändert
1111	(12683)	-	-	-	-	-	-	-	-	-	-	-	0.00 %
0110	(9089)	624	312	112	-	-	-	-	-	-	-	-	11.53 %
1110	(9069)	729	132	-	1462	246	205	84	-	-	-	-	31.51 %
0111	(7380)	239	622	-	1038	-	-	-	334	296	75	-	35.28 %
0010	(6206)	-	-	50	-	-	-	-	-	-	-	-	0.81 %
0100	(5089)	-	-	112	-	-	-	-	-	-	-	-	2.20 %
0011	(5057)	1015	-	78	-	-	-	-	-	-	57	-	22.74 %
1100	(4544)	-	507	83	-	-	-	32	-	-	-	-	13.69 %
1011	(3153)	169	-	-	-	-	223	38	327	-	24	50	26.36 %
0101	(2752)	-	459	15	-	-	-	-	-	-	127	-	21.84 %
1101	(2573)	-	110	-	-	331	-	11	-	242	93	25	31.56 %
0001	(2481)	-	-	35	-	-	-	-	-	-	-	-	1.41 %
1010	(2200)	531	-	16	-	-	-	39	-	-	-	-	26.64 %
0000	(2178)	-	-	-	-	-	-	-	-	-	-	-	0.00 %
1000	(1392)	-	-	103	-	-	-	-	-	-	-	-	7.40 %
1001	(944)	-	-	10	-	-	-	18	-	-	98	-	13.35 %

Tabelle 4.2: Die Tabelle zeigt die Verteilung der Segmentierungen für Anfragen mit fünf Wörtern im Webis Query Segmentation Corpus. Links ist die Segmentierung angegeben, in Klammern findet sich die Anzahl der Segmentierer, die so segmentiert hat. Bei 5-Wort-Anfragen gibt es vier potenzielle Trennstellen, 1 ist gesetzt, wenn dort tatsächlich getrennt wird, 0 wenn nicht. In der Horizontalen ist die Anzahl der Segmentierungen angegeben, zu der gewechselt wird, d.h. die beim Vereinigen durch das Entfernen von Trennstellen entsteht. Die leere Segmentierung 1111 wird beim Vereinigen nicht mit berücksichtigt.

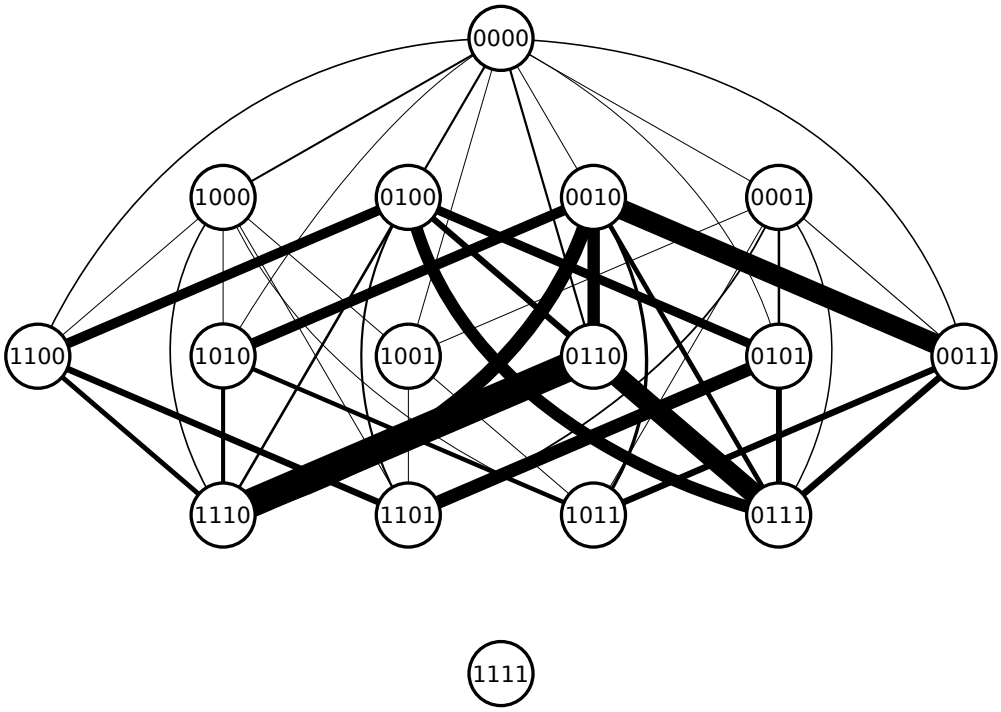

Abbildung 4.2: Die Grafik veranschaulicht das Vereinigen der Segmentierungen der 5-Wort-Anfragen. Die Linienstärke entspricht der Anzahl der gewechselten Segmentierungen. Von der 1110-Segmentierung gibt es die meisten Vereinigungen mit der 0110-Segmentierung, die hinterher die meisten Stimmen besitzt.

Kapitel 5

Zusammenfassung und Ausblick

In dieser Arbeit wurde nach mehrdeutigen Anfragen gesucht, die als Grenzfälle für Segmentierungsalgorithmen von Interesse sind. Damit mehrdeutige Suchanfragen in den Algorithmen zum automatischen Segmentieren besser berücksichtigt werden können, sollten solche Anfragen analysiert und Eigenschaften abgeleitet werden.

Zuerst wurde mittels Korpusanalyse versucht, den Anforderungen entsprechende Suchanfragen im Webis Query Segmentation Corpus zu finden, das von Hagen et al. im Jahr 2012 zur Evaluierung von Segmentierungsalgorithmen erstellt wurde. Dafür wurden über 50 000 Anfragen aus dem 2006 veröffentlichten AOL-Suchlog ausgewählt und jeweils mindestens von zehn Menschen segmentiert. Dadurch, dass die segmentierenden Menschen meistens unterschiedlich Segmentierten, verteilen sich die „Stimmen“ der Segmentierer auf verschiedene Segmentierungen einer Anfrage. Je nach Verteilung können die Anfragen in die Kategorien „sicher“, „mittel“ und „unsicher“ eingeteilt werden. Bei der Kategorie „sicher“ wird eine Segmentierung bevorzugt, während bei „unsicheren“ Anfragen keine Segmentierung erkennbar ist, die häufiger gewählt wird, als andere.

Anhand von Wikipediatiteln wurden die Anfragen im Korpus auf sich überschneidende Konzepte überprüft. Hierbei wurde, nach Aussortierung selten erkannter Wikipediatitel, eine Menge von Anfragen einzeln von Hand auf Mehrdeutigkeit untersucht. Diese umfassten Anfragen mit sich gegenseitig überlappenden Wikipediatiteln sowie solche, bei denen ein langer Wikipediatitel einen kürzeren einschließt.
In Anfragen, in denen ein längerer Wikipediatitel einen kürzeren einschließt, wird in 66,71 % aller Fälle der längere Wikipediatitel gewählt. In Anfragen, in denen sich zwei Wikipediatitel gegenseitig überlappen, entscheiden sich insgesamt 64,45 % der Segmentierer für den Wikipediatitel mit der höheren N-Gramm-Häufigkeit.

Durch manuelle Überprüfung ließen sich dann, ausschließlich in der Menge der Anfragen mit sich gegenseitig überlappenden Wikipediatiteln, nur 25 mehrdeutige Anfragen finden. Diese haben eine Länge von drei oder vier Wörtern und bestehen hauptsächlich aus Strict Noun Phrases (SNP). Die relativ kurze Anfragenlänge könnte der generellen Verteilung im Korpus geschuldet sein, in dem größtenteils 3- und 4-Wort-Anfragen vorhanden sind. Ebenso lässt sich aber auch vermuten, dass Nicht-SNPs, aus denen längere Suchanfragen meist bestehen, seltener Mehrdeutigkeiten enthalten.
Es bleibt festzuhalten, dass sehr wenige mehrdeutige Anfragen, nur ca. 0,05 %, im Webis Query Segmentation Corpus gefunden werden konnten. Allerdings fanden sich wie angenommen die meisten davon unter den als „unsicher“ kategorisierten Anfragen mit überlappenden Wikipediatiteln. Jedoch können nur in wenigen Fällen die durch Wikipediatitel repräsentierten überlappenden Konzepte klar als Ursache für die Unsicherheit der Segmentierer ermittelt werden.

Der vorgeschlagene Algorithmus zum Vereinigen ähnlicher Segmentierungen nimmt an, dass sich zwei Segmentierungen ähneln, wenn durch das Weglassen einer bestimmten Anzahl von Trennstellen die eine Segmentierung aus der anderen erzeugt werden kann. Bei der Betrachtung von 5-Wort-Anfragen wird gezeigt, dass die Segmentierung mit einem 2-Wort-Segment am Anfang und am Ende (0110) nach der leeren Segmentierung die beliebteste Form der Segmentierung ist. Durch das Vereinigen werden viele Segmentierungen mit drei Segmenten zu 2-Wort-Segmenten. Jedoch bleiben Segmentierungen mit drei Segmenten mit 32,86 % die häufigste Form, gefolgt von Segmentierungen mit zwei Segmenten, die 27,36 % ausmachen.

Die Anzahl der in dieser Arbeit gefundenen mehrdeutigen Anfragen ist sehr klein. Um allgemeinere Merkmale und Eigenschaften ermitteln zu können braucht man eine größere Zusammenstellung dieser Anfragen. Die Analyse von Suchanfragen mit überlappenden Wikipediatiteln im gesamten AOL-Suchlog könnte hier noch mehr Ergebnisse bringen. Es könnte sich auch eine genauere Betrachtung von Anfragen mit mehr als zwei überlappenden Wikipediatiteln lohnen.
Denkbar wäre auch eine Methode zum Vereinigen ähnlicher Segmentierungen, in der die Segmentierung von Wikipediatiteln als Kriterium verwendet wird.

Anhang A

Tabellen

	Wörter								
Verteilung	3	4	5	6	7	8	9	10	Summe
10	1804	679	185	56	23	15	7	5	2774
9,1	2666	1120	418	166	95	36	21	11	4533
8,2	1860	847	289	94	42	23	13	7	3175
8,1,1	1412	762	422	207	91	48	27	15	2984
7,3	1648	745	259	71	20	15	6	4	2768
7,2,1	2008	1075	493	217	89	46	18	4	3950
7,1,1,1	257	345	224	134	74	49	30	14	1127
6,3,1	1776	954	447	153	67	38	19	7	3461
6,2,2	726	317	139	47	25	7	3	-	1264
6,2,1,1	680	681	423	236	94	46	22	15	2197
6,1,1,1,1	-	102	101	87	51	22	18	10	391
5,1,1,1,1,1	-	26	38	34	22	15	12	6	153
	14 837	7653	3438	1502	693	360	196	98	28 777

Tabelle A.1: Die Tabelle zeigt die Verteilung der „sicheren“ Anfragen im Webis Query Segmentation Corpus, die mit 53,86 % (28 777 von 53 432) den größten Anteil aller vorhandenen Anfragen ausmachen.

	Wörter								
Verteilung	3	4	5	6	7	8	9	10	Summe
6,4	1596	676	200	68	20	5	6	1	2572
5,3,2	1270	542	252	81	33	11	5	5	2199
5,3,1,1	626	556	386	197	64	34	16	16	1895
5,2,2,1	514	448	285	122	58	25	10	5	1467
5,2,1,1,1	-	263	251	184	96	40	23	15	872
4,1,1,1,1,1,1	-	4	19	16	28	10	1	2	80
	4006	2489	1393	668	299	125	61	44	9085

Tabelle A.2: Anfragenverteilung in der Kategorie „mittel“ im Webis Query Segmentation Corpus, die insgesamt 17 % des Korpus ausmacht (9085 von 53 432).

	Wörter								
Verteilung	3	4	5	6	7	8	9	10	Summe
5,5	567	309	112	35	15	2	3	-	1043
5,4,1	1607	865	356	141	44	27	13	4	3057
4,4,2	582	267	98	38	9	5	1	-	1000
4,4,1,1	287	292	167	80	40	17	6	2	891
4,3,3	479	198	83	37	12	6	2	-	817
4,3,2,1	954	788	455	200	79	24	19	9	2528
4,3,1,1,1	-	262	253	146	68	41	20	8	798
4,2,2,2	171	119	46	23	10	2	-	3	374
4,2,2,1,1	-	292	242	157	74	35	11	2	813
4,2,1,1,1,1	-	77	140	115	74	55	29	13	503
3,3,3,1	133	113	49	35	14	5	1	1	351
3,3,2,2	207	155	66	35	14	6	3	1	487
3,3,2,1,1	-	288	207	120	52	39	15	3	724
3,3,1,1,1,1	-	44	70	48	34	22	9	3	230
3,2,2,2,1	-	129	110	62	32	16	3	2	354
3,2,2,1,1,1	-	147	202	163	100	43	26	13	694
3,2,1,1,1,1,1	-	26	71	88	84	45	18	17	349
3,1,1,1,1,1,1,1	-	-	10	9	14	11	4	5	53
2,2,2,2,2	-	6	6	1	2	-	1	-	16
2,2,2,2,1,1	-	36	42	20	15	7	4	5	129
2,2,2,1,1,1,1	-	14	44	56	41	26	9	10	200
2,2,1,1,1,1,1,1	-	1	17	21	32	18	12	9	110
2,1,1,1,1,1,1,1,1	-	-	2	4	11	5	13	8	43
1,1,1,1,1,1,1,1,1,1	-	-	-	-	-	4	2	-	6
	4987	4428	2848	1634	870	461	224	118	15 570

Tabelle A.3: Die Segmentierungen in der Kategorie „unsicher“ im Webis Query Segmentation Corpus. Das Korpus besteht zu 29,14 % aus solchen Anfragen (15 570 von 53 432).

Verteilung	3 Wörter	4 Wörter	5 Wörter	6 Wörter	7 Wörter	8 Wörter	9 Wörter	10 Wörter	Summe
sicher	0	37	76	84	55	45	22	11	330
mittel	0	15	22	24	19	8	6	3	97
unsicher	240	143	142	75	37	25	8	7	677
verbleibend	208	415	321	209	189	116	56	26	1540
	448	610	561	392	300	194	92	47	2644

Tabelle A.4: Die Tabelle zeigt, in welchen Anfragen der Kategorie „unsicher“ in denen ein Wikipediatitel einen anderen beinhaltet, sicher, mittel oder unsicher segmentiert wurden. Wurden beide Wikipediatitel zusammen nicht mindestens sieben Mal segmentiert, ist die Anfrage als „verbleibend“ gezählt.

Verteilung	3 Wörter	4 Wörter	5 Wörter	6 Wörter	7 Wörter	8 Wörter	9 Wörter	10 Wörter	Summe
sicher	0	14	47	40	30	28	6	12	177
mittel	0	9	13	5	11	5	4	1	48
unsicher	61	79	43	36	20	9	2	2	252
verbleibend	111	332	239	148	173	95	40	13	1151
	172	434	342	229	234	137	52	28	1628

Tabelle A.5: Die Tabelle zeigt, in welchen Anfragen der Kategorie „unsicher“ sich gegenseitig überlappende Wikipediatitel sicher, mittel oder unsicher segmentiert wurden. Wurden beide Wikipediatitel zusammen nicht mindestens sieben Mal segmentiert, ist die Anfrage als „verbleibend“ gezählt.

Verteilung	3 Wörter	4 Wörter	5 Wörter	6 Wörter	7 Wörter	8 Wörter	9 Wörter	10 Wörter	Summe
sicher	224	297	293	207	96	46	29	14	1206
mittel	28	22	8	5	2	0	1	0	66
unsicher	0	3	2	4	0	1	0	0	10
verbleibend	74	152	150	102	64	54	56	14	666
	326	474	453	318	162	101	86	28	1948

Tabelle A.6: In der Tabelle ist zu sehen, in welchen Anfragen der Kategorie „sicher“ sich gegenseitig überlappende Wikipediatitel sicher, mittel oder unsicher segmentiert wurden. Wurden beide Wikipediatitel zusammen nicht mindestens sieben Mal segmentiert, ist die Anfrage als „verbleibend“ gezählt.

Verteilung	3 Wörter	4 Wörter	5 Wörter	6 Wörter	7 Wörter	8 Wörter	9 Wörter	10 Wörter	Summe
sicher	549	522	478	293	127	63	39	14	2085
mittel	114	54	14	10	4	0	1	0	197
unsicher	0	2	4	3	0	0	2	0	11
verbleibend	53	121	111	64	51	35	36	15	486
	716	699	607	370	182	98	78	29	2779

Tabelle A.7: In der Tabelle kann abgelesen werden, in welchen Anfragen der Kategorie „sicher“ ineinander liegende Wikipediatitel sicher, mittel oder unsicher segmentiert wurden. Wurden beide Wikipediatitel zusammen nicht mindestens sieben Mal segmentiert, ist die Anfrage als „verbleibend“ gezählt.

Literaturverzeichnis

[BCS09] Michael Bendersky, W. Bruce Croft und David A. Smith. Two-stage query segmentation for information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seiten 810–811, New York, USA, 2009. ACM.

[BCS10] Michael Bendersky, W. Bruce Croft und David A. Smith. Structural annotation of search queries using pseudo-relevance feedback. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, Seiten 1537–1540, New York, USA, 2010. ACM.

[BCS11] Michael Bendersky, W. Bruce Croft und David A. Smith. Joint annotation of search queries. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Seiten 102–111, Stroudsburg, USA, 2011. ACL.

[BGAG10] David J. Brenes, Daniel Gayo-Avello und Rodrigo Garcia. On the fly query entity decomposition using snippets. In *Proceedings of the 1st Spanish Conference on Information Retrieval*, CERI, 2010.

[BW07] Shane Bergsma und Qin Iris Wang. Learning noun phrase query segmentation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Seiten 819–826, Prague, Czech Republic, 2007. ACL.

[DDQ+13] Xiao Ding, Zhicheng Dou, Bing Qin, Ting Liu und Ji-Rong Wen. Improving web search ranking by incorporating structured annotation of queries. In *EMNLP*, Seiten 468–478, Seattle, USA, 2013. ACL.

[GXLC08] Jiafeng Guo, Gu Xu, Hang Li und Xueqi Cheng. A unified and discriminative model for query refinement. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seiten 379–386, New York, USA, 2008. ACM.

[HGM+10] Jian Huang, Jianfeng Gao, Jiangbo Miao, Xiaolong Li, Kuansan Wang, Fritz Behr und C. Lee Giles. Exploring web scale language models for search query processing. In *Proceedings of the 19th International Conference on World Wide Web*, Seiten 451–460, New York, USA, 2010. ACM.

[HPBS12] Matthias Hagen, Martin Potthast, Anna Beyer und Benno Stein. Towards optimum query segmentation: In doubt without. In *21st ACM International Conference on Information and Knowledge Management*, Seiten 1015–1024. ACM, 2012.

[HPSB10] Matthias Hagen, Martin Potthast, Benno Stein und Christof Bräutigam. The power of naïve query segmentation. In *33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seiten 797–798. ACM, 2010.

[HPSB11] Matthias Hagen, Martin Potthast, Benno Stein und Christof Bräutigam. Query segmentation revisited. In *20th International Conference on World Wide Web*, Seiten 97–106. ACM, 2011.

[JRMG06] Rosie Jones, Benjamin Rey, Omid Madani und Wiley Greiner. Generating query substitutions. In *Proceedings of the 15th International Conference on World Wide Web*, Seiten 387–396, New York, USA, 2006. ACM.

[KGA+10] Julia Kiseleva, Qi Guo, Eugene Agichtein, Daniel Billsus und Wei Chai. Unsupervised query segmentation using click data: Preliminary results. In *Proceedings of the 19th International Conference on World Wide Web*, Seiten 1131–1132, New York, USA, 2010. ACM.

[LHZW11] Yanen Li, Bo-Jun Paul Hsu, ChengXiang Zhai und Kuansan Wang. Unsupervised query segmentation using clickthrough for information retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seiten 285–294, New York, USA, 2011. ACM.

[MSRG+11] Nikita Mishra, Rishiraj Saha Roy, Niloy Ganguly, Srivatsan Laxman und Monojit Choudhury. Unsupervised query segmentation using only query logs. In *Proceedings of the 20th International Conference Companion on World Wide Web*, Seiten 91–92, New York, USA, 2011. ACM.

[RCBR13] Rohan Ramanath, Monojit Choudhury, Kalika Bali und Rishiraj Saha Roy. Crowd prefers the middle path: A new IAA metric for crowdsourcing reveals turker biases in query segmentation. In *ACL*, Seiten 1713–1722, 2013.

[RMB03] Knut Magne Risvik, Tomasz Mikolajewski und Peter Boros. Query segmentation for web search. In *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary, May 2003.

[SRGCL12] Rishiraj Saha Roy, Niloy Ganguly, Monojit Choudhury und Srivatsan Laxman. An IR-based evaluation framework for web search query segmentation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seiten 881–890, New York, USA, 2012. ACM.

[TP08] Bin Tan und Fuchun Peng. Unsupervised query segmentation using generative language models and wikipedia. In *Proceedings of the 17th International Conference on World Wide Web*, Seiten 347–356, New York, USA, 2008. ACM.

[WHLC13] Haocheng Wu, Yunhua Hu, Hang Li und Enhong Chen. Query segmentation for relevance ranking in web search. *CoRR*, abs/1312.0182, 2013.

[WM07] Ryen W. White und Dan Morris. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seiten 255–262, New York, USA, 2007. ACM.

[YS09] Xiaohui Yu und Huxia Shi. Query segmentation using conditional random fields. In *Proceedings of the 1st International Workshop on Keyword Search on Structured Data*, Seiten 21–26, New York, USA, 2009. ACM.

[ZCL+13] Wei Zhang, Yunbo Cao, Chin-Yew Lin, Jian Su und Chew-Lim Tan. Learning a replacement model for query segmentation with consistency in search logs. In *6th International Joint Conference on Natural Language Processing*, Seiten 28–36, Nagoya, Japan, 2013. ACL.

[ZSH+09] Chao Zhang, Nan Sun, Xia Hu, Tingzhu Huang und Tat-Seng Chua. Query segmentation based on eigenspace similarity. In *Proceedings of the 4th International Joint Conference on Natural Language Processing*, Seiten 185–188, Stroudsburg, USA, 2009. ACL.