Andrea Stubbe

Klassifikation von Texten nach Genre

Magisterarbeit
im Fach Computerlinguistik
LMU München

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Formelverzeichnis

1 Einleitung

Diese Arbeit stellt ein System zur automatischen Erkennung des Genres eines Textes vor. »Genre« meint hier eine Klasse von Dokumenten ähnlicher Form und Funktion und stellt, neben Thema oder Inhalt, eine zusätzliche Dimension bei der Beschreibung von Texten dar. In einem ersten Schritt wurde eine Genre-Hierarchie ausgearbeitet und ihre einzelnen Elemente definiert. Anschließend wurden die entstandenen Klassen und ihre Besonderheiten untersucht und diese Merkmale mit Hilfe von Perl-Programmen aus den Texten extrahiert. Aus den Ergebnissen dieser Analyse entstand für jedes Genre ein Klassifikator der bestimmt, ob ein Text zu dieser Gruppe gehört oder nicht. Um deren Qualität zu verbessern und teils auch eine eindeutige Zuordnung in ein Genre zu erreichen, wurden verschiedene Verfahren zur Kombination der einzelnen Klassifikatoren entwickelt. Zusätzlich wurden diverse Algorithmen aus dem Bereich Knowledge-Discovery zur Erkennung verwendet.

Die Evaluation ergibt, dass das von mir entwickelte Verfahren, bei dem für jedes Genre anhand seiner spezifischen Merkmale ein Klassifikator erstellt wird, die besten Ergebnisse liefert. Die durchschnittlichen Werte für Recall und Precision liegen bei knapp 60% und 75%, wobei es jedoch starke Unterschiede zwischen den einzelnen Genres gibt.

Anwendungsgebiete

Die Berücksichtigung des Genres führt in vielen Bereichen der Computerlinguistik zu Verbesserungen. Bei Suche und Information Retrieval können die Anfragen durch Angabe des gewünschten (oder Auschluss von ungewünschten) Genres präzisiert werden [DEW, ROU]. Man kann gezielt entweder Gedichte oder wissenschafliche Abhandlungen über Rosengärten suchen, oder sich, wenn man aktuelle Informationen finden möchte, auf journalistische Texte beschränken. Auch auf Spezialbereiche, wie die Suche nach Bildern, könnte sich dies vorteilhaft auswirken, da bestimmte Arten von Bildern in manchen Genres verstärkt verwendet werden. Möchte man ein Foto von einer Person finden, so liegt es nahe, zunächst in Porträts oder Biografien zu recherchieren.

Die Berücksichtigung der *Form* bestimmter Textarten verbessert die Qualität der Treffer noch weiter: bei der Suche in FAQs kann zum Beispiel verlangt werden, dass alle Begriffe im gleichen Frage-Antwort-Paar vorkommen [CRK]. Auch Algorithmen zur automatischen Zusammenfassung profitieren von den Form-Eigenheiten, da je nach Genre die wichtigen Informationen an verschiedenen Stellen stehen. Nachrichten fassen den Inhalt in den ersten paar Sätzen zusammen, Romane oder Erzählungen folgen dem typischen Spannungsbogen aus Einleitung, Hauptteil (mit Höhepunkt der Handlung) und Schluss.

Der statistische Fehler bei der Generalisierung von Ergebnissen wird geringer, wenn die Trainings-Elemente möglichst gut die gewünschte Klasse repräsentieren und keine unpassenden Daten enthalten. Deswegen führt auch die Beschränkung auf das selbe Genre in Training und Test zu Verbesserungen. So konnte gezeigt werden, dass durch diese Methode die Performance eines probabilistischen Parsers steigt [ILL]. Bestimmte objektlose Konstruktionen kommen beispielsweise im Englischen nur in Rezepten vor, in anderen Genres könnte diese Alternative beim Parsing somit ausgeschlossen werden. Ähnliches gilt beim Tagging und der Word-Sense-Disambiguierung, da bestimmte Bedeutungen von Wörtern in manchen Genres häufiger als in anderen auftreten. So

wird »pretty« in informellen Texten häufiger im Sinne von »ziemlich« statt »hübsch« verwendet, das Wort »trend« ist 35 mal häufiger ein Verb in Nachrichten (im *Journal for Commerce*) als in wissenschaftlichen Artikeln (in *Sociological Abstracts*) [KES]. Mit diesne Erkenntnissen können Übersetzungsprogramme und andere von diesen Technologien abhängige Anwendungen optimiert werden.

Auch beim Kompilieren von Korpora, Wortlisten und ähnlichen Ressourcen für linguistische Arbeiten kann das Auswählen bzw. Filtern von Dokumenten aus bestimmten Genres das Ergebnis verbessern. Möchte man zum Beispiel ein Lexikon zur Fehlerkorrektur erstellen, so könnte man ausschließlich auf professionelle Texte zurückgreifen.

Gliederung

Als erstes wird der Begriff »Genre« definiert und die Genre-Hierarchie vorgestellt. Es folgt eine Beschreibung der Arbeitsweise bei der Erstellung der Klassifikatoren und eine genauere Betrachtung der Merkmale jedes Genres. Im fünften und sechsten Kapitel werden Klassifikations-Algorithmen und verschiedene Kombinationsverfahren vorgestellt und evaluiert. Den Abschluss bildet ein Ausblick auf mögliche Weiterentwicklungen und Verbesserungen.

Im Anhang sind die Ergebnisse der Evaluation detailliert nach Genre aufgeschlüsselt beigefügt. Die Programme, der Trainings- und Testkorpus sowie diverse benötigte Wortlisten finden sich auf beigelegter CD, sowie teilweise unter *http://www.astro-susi.de/genre/*.

Die Rechte an den Texten im Korpus liegen bei den jeweiligen Autoren. Die Wortlisten wurden von mir zusammengestellt und dürfen frei verwendet werden. Eine Ausnahme bilden die Liste der Vornamen und die der 200 000 häufigsten englischen Wörter; diese stammen vom Centrum für Informations- und Sprachverarbeitung (CIS) der LMU München.

2 Genre

Das Genre eines Textes wird im wesentlichen durch *Struktur*, *Schreibstil* und kommunkative *Funktion* bestimmmt [DEWE]. Die beiden ersten Begriffe werden oft unter *Form* zusammengefasst. Texte eines Genres haben bestimmte Merkmale gemeinsam, durch die sie sich von Dokumenten anderer Klassen abgrenzen lassen [DEWE].

Form und Funktion

Die *Struktur* beschreibt, ob und wie der Text gegliedert ist, also ob er Listen, Überschriften, Absätze, Tabellen, Hervorhebungen und andere Layoutmerkmale enthält. Der *Schreibstil* wird durch das verwendete Vokabular, den Satzbau und die Diskursstrategie (also die Folge von Behauptungen, Aussagen, Fragen usw.) bestimmt. Für Whitelaw und Argamon ist der Stil eine Wahl des Autors aus mehreren Möglichkeiten auf jeder dieser drei Ebenen. Er hängt nicht nur vom Genre ab, sondern auch von sozialem Status, Herkunft und Persönlichkeit des Autors und seiner Leserschaft [WHI]. Plum und Cowling konnten einen Zusammenhang zwischen verwendeter Zeitform und gesellschafltlicher Klassenzugehörigkeit des Verfassers herstellen [PLU], Argamon et. al. entdeckten Unterschiede im Schreibstil von Männern und Frauen [ARG].

Die kommunikative *Funktion* beschreibt Zweck und Bedeutung eines Textes: Soll der Leser informiert oder überzeugt werden? Dient der Text der Zerstreuung und Unterhaltung oder legt er Regeln und Gesetze fest? Dieser Aspekt beschreibt die semantische Komponente eines Textes.

Oft besteht ein Zusammenhang zwischen Form und Funktion, da sich bestimmte Formatierungen (*Signalling Devices*) als besonders geeignet erwiesen haben, die Erfassung der Bedeutung eines Textes zu unterstützen. Durch die allgemeine Verwendung dieser Form funktioneren diese Merkmale beim Leser als ein Auslöser, um das Genre eines Textes zu erkennen. Die Form erleichtert also das Erfassen des Zwecks eines Textes; weicht sie stark von der Norm ab erschwert dies das Textverständnis erheblich – man stelle sich ein Lexikon vor, das weder alphabetisch sortiert ist, noch irgendeine Art der Hervorhebung für die erklärten Begriffe verwendet. Die Aussage der *Signalling Devices* steht hier im Widerspruch zu den Erwartungen des Lesers, wodurch zusätzliche kognitive Ansprüche an ihn gestellt werden. [TOC]

Einige Autoren schließen noch andere Merkmale als Form und Funktion ein. Für [TOC] und [ROU] spielt das physische Medium (Buch, Broschüre) bzw. das Interface eine Rolle, Probanden bei [DEWE] wählten unter anderem Verfasser, Qualität oder die Verfügbarkeit von Dokumenten. Da diese Eigenschaften nicht automatisch erkannt werden können und außerdem eher eigene Dimensionen darstellen als das Genre defineren, werden diese im folgenden nicht betrachtet.

Auch das Themengebiet eines Textes hat mit seinem Genre zunächst nichts zu tun, teilweise wird sogar behauptet, diese Eigenschaften seien orthogonal [FIN]. Wie [KAC] jedoch richtig bemerkt haben, besteht durchaus ein Zusammenhang zwischen Inhalt und Art eines Textes. Elfen und Feen sind zum Beispiel ähnlich selten Gegenstand von wissenschaftlichen Abhandlungen, wie Gedichte über Neurologie geschrieben werden. In Abschnitt 2.1 kann man sehen, dass manche Arbeiten Thema und Genre teilweise auch vermischen.

Hierarchien

Die verschiedenen Genres lassen sich in eine Hierarchie einordnen. Diejenigen innerhalb der selben Haupt-Kategorie haben einige Eigenschaften in Form oder Funktion gemeinsam und unterscheiden sich in anderen. Eine solche Gliederung ist hilfreich bei vielen Anwendungen, ganz besonders bei der Suche. Man kann sich beispielsweise allgemein auf journalistische Texte beschränken, wenn man aktuelle Informationen sucht, ohne sich über die Unterschiede der einzelnen Journalismus-Genres im Klaren zu sein. Außerdem kann man den Hierarchie-Baum auf der Suche nach bestimmten Textarten durchwandern. [CRK]

Zusätzlich zu dieser *ist-ein*-Hierarchie (Hyponymie) gibt es noch eine *ist-enthalten-in*-Ordnung (Meronymie) [CRW]. Beispiele sind wissenschaftliche Texte, die Code-Listings oder Literaturlisten enthalten, oder Veranstaltungsinformationen mit einem Terminplan. Dieses Enthaltensein kann man ausweiten, wenn man die Texte in einem größeren Kontext sieht: Nachrichten sind Teil einer Zeitung, Kontaktfomulare Teil einer Firmen-Präsentations-Webseite. Rehm definiert einzelne Genres sogar unter anderem dadurch, welche anderen Genres sie enthalten müssen oder können [REH].

2.1 Bisherige Genre-Systeme

2.1.1 Vorhandene Korpora: Brown und Wallstreet-Journal

Der Brown-Korpus war die erste in Genres eingeteilte Textsammlung. Er besteht aus 500 Texten von 1961 mit jeweils etwa 2000 Wörtern und wurde aus 15 verschiedenen Kategorien zusammengestellt, um möglichst repräsentativ zu sein. Auf Grund seiner öffentlichen Verfügbarkeit wird er auch heute noch verwendet (z.B. von [KAC, KES]) oder dient als Grundlage für andere Korpora wie den schwedischen SUC (Stockholm-Umeå-Corpus [WAS]). Allerdings wurde dieser Korpus nicht ursprünglich zur Genre-Erkennung erstellt und weist deswegen einige Mängel auf. So sind manche der Genres zu allgemein (»General Fiction« oder »Miscellaneous«), andere zu speziell oder sogar inhaltsbezogen (»Religion«) [STA]. Manche Kategorien fehlen auch ganz, beispielsweise Listen, Lexika oder Interviews sowie alle neueren und Internet-spezifischen Genres.

A	Presse: Reportage
B	Presse: Editorials (inkl. Leserbriefe)
C	Presse: Reviews
D	Religion
E	Skills and Hobbies
F	Populare Lore (Populärwissenschaftliches)
G	Belles Lettres etc. (Briefe, Biographien, Memoiren)
H	Miscellaneous (Regierung, Industrie, Universitätsdokumente)
J	Learned
K	Fiction: General
L	Fiction: Mistery
M	Fiction: Science Fiction
N	Fiction: Adventure
P	Fiction: Romance
R	Humor

Tabelle 2.1: Genres im Brown-Korpus [KAC]

Da sie mit dem Brown-Korpus unzufrieden waren, teilten Stamatatos et. al. [STA] für ihre Versuche den Wallstreet-Journal-Korpus in diverse journalistische Genres auf. Die Einteilung ist allerdings nicht vollständig – es fehlen zum Beispiel Interview und Rezension. Die vier vorhandenen Klassen sind jedoch themenunabhängig und klar voneinander abgrenzbar und eignen sich daher gut für Versuche zur automatischen Klassifizierung.

1 Editorials
2 Reportage
3 Leserbriefe
4 Nachrichten

Tabelle 2.2: Genres im Wall Street Journal [STA]

2.1.2 Neue Klassifikationen

Dewe, Karlgren und Bretan [DEWE] entwickelten 1998 mit Hilfe von Umfragen ein eigenes System von Genres. Wichtig war ihnen dabei, dass die Genres den Erwartungen der Benutzer entsprechen und zugleich einfach modellier- und identifizierbar sind. Dazu wurde Studenten und Wissenschaftlern die Frage »Welche Genres gibt es im WWW?« gestellt. Die 69 Antworten zeigen, dass das Verständnis von »Genre« nicht einheitlich ist. Einige der Interviewten vermischen Gerne und Inhalt (Sport, Pornographie) oder nehmen die Absicht des Autors (»I guess we have to be on the net too«), den Verfasser (non-governmental organization info), Qualität (langweilige Homepages) und das Umfeld der Veröffentlichung (intern/öffentlich) als Merkmale. Die verwertbaren dieser Antworten wurden von den Autoren zusammengefasst. Das Ergebnis ist eine gut durchdachte und vor allem recht knappe Klassifizierung, die auch als Grundlage meiner eigenen Einteilung dient:

1 Private Homepages
2 Öffentliche, kommerzielle Homepages
3 Interaktive Seiten (mit Feedback: Dialoge, durchsuchbare Indizes)
4 Journalistische Texte (Nachrichten, Editorials, Rezensionen, Reportagen, E-Zines)
5 Berichte (wissenschaftliche Texte, Gesetze und öffentliche Materialien, formale Texte)
6 Anderer Fließtext
7 FAQ
8 Linklisten
9 Andere Listen und Tabellen
10 Diskussionen (Forum, Usenet)
11 Fehlermeldungen

Tabelle 2.3: Dewes Genre-Einteilung [DEWE]

Diese Einteilung wurde zur Evaluierung den befragten Personen vorgelegt, die größtenteils einverstanden damit waren. Kritisiert wurde unter anderem das Genre »Interaktive Seiten« und dass »anderer Fließtext« zu weit gefasst sei.

Crowston und Williams [CRW96] analysierten 100 Webseiten und entdeckten dabei 48 verschiedene Kategorien. Ihr Ziel war es, neue Genres im WWW zu identifizieren. Die hohe Zahl trotz der geringen Anzahl untersuchter Seiten erklärt sich dadurch, dass die einzelnen Genres sehr eng gefasst sind und teilweise den Inhalt statt Form und Funktion beschreiben. So werden zum Beispiel Filmographie und Diskographie als unterschiedliche Genres identifiziert.

Auch Roussinov et. al. [ROU] haben 2001 versucht, eine eigene Genre-Hierarchie zu schaffen. Wie [DEWE] hatten sie dabei den praktischen Nutzen, speziell für die Suche im Web, und die maschinelle Erkennbarkeit im Sinn. Sie verwendeten ebenfalls User-Interviews, in denen sie 184 Menschen baten, das Ziel ihrer Recherchen im Internet zu beschreiben und dafür hilfreiche Genres zu benennen. Als Grundlage diente die umfangreiche Liste von [CRW], der sie noch 44 neue Genres hinzufügten – darunter das beliebte »Sonstige«, das hier unter anderem die Subgenres Kalender, Kontakt, Fahrplan, Forum oder Aktienkurse enthält. Das Ergebnis dieser Prozedur ist leider nicht mehr unter der im Paper angegebenen URL erreichbar. Anschließend wurden intuitiv fünf Gruppen identifiziert, die zur Erfüllung der Suchwünsche geeignet sind und alle Genres, die mit einer geschätzten (!) Genauigkeit von mindestens 60% erkannt werden, darin eingeordnet.

1 Topics: Homepage (geschäftlich, Berühmtheiten), Orte, »special topics«
2 Publikationen: Artikel, News-Bulletins
3 Produkte: Produktinformation und -listen, Werbung, Bewertungen, Bestellformulare
4 Bildung: Glossar, Kursliste, Anleitungen
5 FAQ

Tabelle 2.4: Genres von Roussinov [ROU]

Wie man sieht, ist diese Liste sehr auf spezielle Suchprobleme fokussiert, wozu sie auch ganz geeignet scheint. Allerdings fehlen sehr viele Textarten.

Einen anderen Ansatz verfolgen [KES], die Genre nicht als eine einzelne Eigenschaft von Texten sehen, sondern als eine multidimensionale Kombination mehrerer Merkmale oder Facetten. Eine Kategorie entspricht hier einer spezifischen Zusammensetzung; beispielsweise könnte ein Kommentar als analysierend + öffentlich + Non-Fiction + mittleres Niveau klassifiziert werden. Der Vorteil ist, dass das System durch neue Verknüpfungen beliebig erweiterbar ist und dadurch auch bisher unbekannte Kategorien erfassen kann. Ein Argument gegen Einteilungen dieser Art ist, dass die so entstandenen Kategorien, obwohl sie empirisch erfasst werden können, keinen eigentlichen *Sinn* haben [KAC]. Dadurch können sie den Benutzer bei Suche, Retrieval und Ähnlichem auch nicht unterstützen.

2.1.3 Web-Genres

Mit dem neuen Medium Internet entstehen neuen Genres. Diese sind entweder Erweiterungen schon existierender Genres oder bis dahin unbekannte Neuerfindungen. Für Shepherd und Watters lassen sie sich, außer durch Inhalt und Form, durch das zusätzliche Attribut *Funktionalität* beschreiben. Dieses wird eher vage als die durch das Medium neu verfügbaren »Möglichkeiten« definiert [SHE98]. Betrachtet man Genre allerdings ohnehin schon als Texte ähnlicher Form und Funktion (statt Inhalt), so stellt man fest, dass diese neue Facette ohne Probleme der Funktion zugeordnet werden kann. Die neue Möglichkeit der Interaktion, um ein typisches Beispiel zu wählen, ist nichts anderes als die kommunkative Funktion der Aufforderung an den Benutzer, zu Handeln: durch Eingabe von Begriffen bei Suchmaschinen, Auswahl der Sprache etc.

Erweiterungen von Genres um neue medienspezifische Elemente führen auch nicht automatisch zu neuen Genres. Ein Katalog ist auch im Internet eine Auflistung von Waren mit Preisangaben und der Information, wo diese Dinge erworben werden können. Ob für den Einkauf der Besuch eines Ladens, eine Telefonhotline oder ein Online-Bestellformular verwendet werden, ändert weder

Funktion noch Struktur und Stil dieser Textart. Viele der von anderen (z.B. [SHE99, CRW]) entdeckten »Cybergenres« sind demnach nur Variationen von altbekannten Genres.

Oft wird auch das Genre »Persönliche Homepage« genannt, das die kommunikative Funktion der Selbstdarstellung im Internet hat [FUR]. Zu häufig zitierten Anhängern dieser These gehören Dillon und Gushrowski, die sogar eine spezifische *Form* ausmachen konnten [DIL]. Allerdings sind viele der genannten Merkmale trivial, z.B. findet man Titel, Bilder, Mailadresse und externe Links auf den meisten Seiten; der Vergleich mit anderen Webseiten fehlt. Außerdem grenzen die Autoren persönliche Homepages von kommerziellen Präsentations-Webseiten nur dadurch ab, dass der Verfasser eine Privatperson ist. Da meiner Auffassung nach der Verfasser für das Genre keine Rolle spielt, können persönliche und geschäftliche Präsentationsseiten zusammengefasst werden (siehe 2.2.1).

Am Beispiel der Homepages kann man auch die zeitliche Begrenztheit von Genres im Web erkennen, da heute eine Vielzahl der persönlichen Webseiten in Form eines Weblogs verfasst werden. Weil die entsprechende Software dafür kostenlos verfügbar und leicht zu bedienen ist und das Ergebnis meist auch noch besser aussieht als selbst erstellte Seiten, wird die Verbreitung vermutlich noch zunehmen. Die klassische Homepage ist hingegen inzwischen eher selten zu finden.

Aber auch wenn viele der identifizierten Cybergenres keine neuen Genres sind, lässt sich nicht abstreiten, dass zusätzliche Textarten im Internet entstanden sind. Eine davon ist das Forum, das zahlreichen Benutzern die Möglichkeit gibt, den Inhalt eines Dokuments zu ergänzen und zu verändern, wodurch eine Kommunikation zwischen den Personen stattfindet. Im Prinzip ist diese Textart eine Abbildung von Gesprächen in Textform, was vor der Erfindung des Internets in dieser Unmittelbarkeit technisch ncht möglich war. Nicht ganz so eindeutig »neu« sind die oben erwähnten Weblogs (oder kurz Blogs), die ein öffentlich zugängliches und von Besuchern der Seite kommentierbares Tagebuch darstellen. Da sie in ihrer Form und Funktion jedoch relativ stark von den alten Tagebüchern abweichen, besonders durch ihre Öffentlichkeit, kann man hier schon von einem neuen Genre sprechen. Ein triviales im Internet aufgetauchtes Genre sind Fehlermeldungen und leere Seiten – wer hätte früher schon leere Blätter und Fehldrucke publiziert?

Beide Genres werden allgemein von den Benutzern akzeptiert, was für Crowston und Williams [CRW98] wichtig ist, um in das bestehende Genrerepertoir aufgenommen zu werden. Ein weiteres Beispiel für ein neu entstandenens Genre sind FAQ-Seiten, die ihren Ursprung im Usenet haben und heute sehr weit verbreitet sind.

Viele Texte im Internet lehnen sich in ihrem Erscheinungsbild an herkömmliche Print-Genres an, um die von Toms und Campbell [TOC] beschriebene Erleichterung der Wahrnehmung zu gewährleisten.

2.2 Eine neue Genre-Hierarchie

Genres sollten also nichts außer Form und Funktion der Texte betrachten und außerdem auch hilfreich für verschiedene Aufgaben wie zum Beispiel Suche in bestimmten Textkategorien, sein. Hierzu tragen auch eine verständliche Definition und Bedeutung sowie nicht zuletzt ein passender Name bei. Die einzelnen Klassen sollten nicht zu feingranular sein und hierarchisch organisiert werden, wobei auch beim Aufbau dieser Hierarchie darauf geachtet werden sollte, dass sie nützlich

und logisch ist. Schließlich ist noch wichtig, dass das System vollständig ist und alle Texte in ein Genre eingeordnet werden können. Es ist möglich, einen Text mehreren Kategorien zuzuordnen.

Aus diesen Anforderungen ergibt sich, dass ein Bottom-Up-Ansatz für diese Aufgabe sinnvoll ist, der eine möglichst große Zahl von Texten empirisch untersucht und dabei stets die Bedürfnisse der Benutzer im Auge behält (vgl. [CRK]).

2.2.1 Entstehung

Ausgehend vom System von [DEWE] entwickelte ich eine eigene Genre-Einteilung. Dabei wurden die oben genannten Kritikpunkte berücksichtigt: Die Klasse »anderer Fließtext« wird aufgespaltet in Literaturgenres, Briefe und diverse Wissen vermittelnde Texte; interaktive Seiten werden zusammen mit Diskussionen und Briefen dem Bereich *Kommunkation* zugeordnet. Zusätzlich werden private und öffentliche Homepages zu *Präsentation* verschmolzen und »Berichte« in *Informationen* und *Berichte* aufgeteilt. Fehlermeldungen wurden mit leeren Seiten und Framesets zu *Nichts* zusammengefasst, da diese keinen Nutzen für den User bieten. Außerdem wurden die Genres in eine Hierarchie eingeordnet und neue Unterpunkte eingefügt. Das Ergebnis war eine erste grobe Ordnung (vgl. Tabelle 2.5).

Bericht	**Kommunikation**	**Informationen**
wissenschaftlich	Brief	Gesetz
offiziell (Militär, Firmen,...)	Formular	Präsentation
Literatur	Blog	FAQ
Roman	Forum	Statistik
Gedicht	**Verzeichnis**	Erklärung
Kurzgeschichte	Personen	**Journalismus**
Nichts	Dinge	Nachricht
	Links	Kommentar
	Sonstige	Interview

Tabelle 2.5 : Das ursprüngliche Genre-System

Besonders wichtig war mir, keine »Sonstiges«-Kategorie auf oberster Ebene einzuführen, sondern Hauptklassen zu finden, die jeden Text erfassen. Außerdem sollte die Hierarchie nicht zu tief sein, da sonst erstens ein sehr großer Korpus benötigt werden würde und zweitens eine zu detaillierte Aufgliederung bei vielen der oben genannten Aufgaben (vgl. 1) nicht hilfreicher wäre.

Um die Qualität dieses Systems zu überprüfen, wurden einige hundert zufällig gewählte Webseiten sowie die interessanten der in der Literatur genannten Genres darin eingeordnet. Besonders die detaillierte Auflistung von [CRW] war hier sehr hilfreich. Neu gefundenen Kategorien wurden in die Hierachie einsortiert. Bei der Analyse stellte ich fest, dass viele Dokumente eine Sammlung oder Mischung aus mehreren Texten darstellen. Ein typisches Beispiel dafür sind Zeitungen, die Beiträge aus den verschiedenen journalistischen Genres enthalten. Da es für das Zusammenfügen von Texten sehr viele Möglichkeiten gibt, wurde für diese Dokumente die künstliche Klasse »Kombinationen« geschaffen. Spezialfälle von Kombinationen sind Forum und Blog, die auch mehrere Beiträge unterschiedlichster Art enthalten können. Wegen ihrer sehr spezifischen Form und Funktion bleiben diese trotzdem als eigene Genres bestehen. Tabelle 2.6 zeigt die erweiterte Hierarchie.

Bericht	**Kommunikation**	**Informationen**
wissenschaftlich	Brief	Gesetze, Regeln
offiziell (Militär, Firmen,...)	Formular	Rezepte und Anleitungen
Erlebnisbericht, Reportage	Blog	FAQ
Gesprächsprotokoll	Forum	Statistiken
Literatur	**Verzeichnis**	Erklärungen
Roman	Zitatsammlungen	Präsentation
Gedicht, Gebet	Katalog	Code-Listing
Kurzgeschichte	Personen	**Journalismus**
Biographie	Dinge (Bücher, Hotels,...)	Rezension
Witze	Links	Nachricht
Essay	Index, Sitemap	Kommentar
Drehbuch	Wortliste (Lexikon, Wörterbuch)	Interview
Nichts	Sonstige	**Kombinationen**

Tabelle 2.6 : Die verbesserte Genre-Hierarchie

Beim Betrachten der Hauptkategorien fällt auf, dass die Texte in »Bericht« wenig gemeinam haben, weder in Struktur noch Funktion oder Stil - teils sind sie strikt gegliedert und sachlich verfasst (offizieller Bericht), teils unterhaltend und locker geschrieben (Reportage). Außerdem sind einige der Genres sehr fein, können aber trotzdem nirgendwo einsortiert werden, beispielsweise Witze oder Linkliste. Manche lassen sich auch zusammenfassen, wie Roman und Kurzgeschichte (da sie sich hauptsächlich in der Länge unterscheiden), oder erweitern, zum Beispiel Interview zu »Interviews und Diskussionen«. Einige Genres sind falsch einsortiert, Essays und Reportagen sind typische Textarten im Journalismus. Deswegen fand eine abschließende Überarbeitung der Hierarchie statt. Bei Auswahl und Abgrenzung der einzelnen journalistischen Genres stand mir Christoph Koch, Journalist bei der Süddeutschen Zeitung, beratend zur Seite.

2.2.2 Ergebnis

Die folgende Liste zeigt das endgültige Ergebnis zusammen mit einer kurzen Beschreibung der Genres.

A. Journalismus

Bei journalistischen Texten kann man zwei Arten unterscheiden: solche, die die Meinung des Autors zeigen (A.1 bis A.4) und neutrale Texte (A.6 und A.7). Die Aufgabe der journalistischen Texte ist hauptsächlich die Information des Lesers.

A.1 Kommentar

Der Kommentar zeigt die Meinung eines Autors zu einem bestimmten und oft aktuellen Thema auf. Dabei geht es weniger um die Vermittlung von Fakten, sondern um eine Analyse der Ereignisse und das Darstellen von Alternativen und Zukunftsperspektiven. Man unterscheidet den abwägenden Kommentar, der Pro- und Contra-Argumente erörtert, und den einseitigen, der ausschließlich die Meinung des Verfassers zeigt. Hierzu zählt auch das Pamphlet, das besonders bissig oder bösartig geschrieben ist [SCJ]. Ziel des Kommentars ist es, den Leser zu überzeugen.

A.2 Rezension

Wie der Kommentar handelt auch die Rezension von der Meinung des Verfassers zu aktuellen Themen – allerdings sind diese hier meist kultureller Natur, wie beispielsweise Bücher, Ausstellungen oder Theateraufführungen; aber auch Restaurants oder Güter (neue Autos, Mode) werden beschrieben. Sie handelt damit von Dingen oder Ereignissen, die der Leser selbst auch konsumieren kann und dient so als eine Art Ratgeber.

A.3 Porträt (Nachruf, Würdigung)

Eine einzelne Person steht bei dieser Textgattung im Mittelpunkt. Auch hier kommentiert der Autor in einem gewissen Sinne, nämlich die Eigenheiten der Person, aber auch Fakten und Daten über ihr Leben werden wiedergegeben. Der Anlass für ein Porträt kann unter anderem ein runder Geburtstag, ein empfangener Preis oder der Tod der Person sein, ein aktueller Bezug wird hergestellt. Der Autor will mit einem Porträt meistens seine Achtung zum Ausdruck bringen und den Leser mit Hintergrundinformationen versorgen. Autobiografien fallen nicht in diese Genre, sondern zählen als Prosa.

A.4 Glosse (Essay, Polemik, Kolumne)

Anders als die oben genannten meinungsbildenden journalistischen Texte erhebt dieses Genre keinen Anspruch auf Aktualität oder Information. Es handelt sich eher um eine nette kleine Geschichte oder Beobachtung, die lustig oder spöttisch dargestellt wird. Ziel der Glosse ist es, den Leser zu unterhalten. In Zeitungen findet man solche Texte häufig an einem festen Platz (z.B. das »Streiflicht« der Süddeutschen Zeitung), man spricht dann von einer Kolumne. [SJC]

A.5 Interview und Diskussion

Bei einem Interview oder einer Diskussion wird ein Gespräch zwischen mehreren Leuten aufgezeichnet. Die sprechenden Personen werden entweder jeweils am Zeilenanfang genannt oder die verschiedenen Gesprächsbeiträge sind durch feste Formatierungen gekennzeichnet (zum Beispiel die Frage fett, die Antwort normal). Typisch für Diskussionen ist die Teilnahme mehrerer, teilweise auch namentlich nicht genannter Personen; der Verlauf ist relativ frei. Das Interview hingegen ist durch Fragen stark gelenkt. Der Journalist kann durch die Auswahl der Themen und die Art seiner Fragen auch seine eigene Meinung ausdrücken. Meist wird nur eine einzige Person oder Gruppe interviewt.

A.6 Nachrichten, Meldungen, Bericht

Eine Nachricht beschreibt neutral, sachlich und verständlich ein aktuelles Ereignis oder fasst Meinungsäußerungen verschiedener Personen zusammen. Sie kann von der wenigen Zeilen umfassenden Kurzmeldung bis zu längeren Berichten gehen. Im Vordergrund steht die Information der Leser. Wichtige Elemente sind Zeit und Ort der Handlung sowie die Angabe von Quellen. Die Nachricht ist nicht chronologisch geordnet, sondern nach Relevanz: die wesentliche Information (Wer, Was, Wann, Wo, Woher) steht im ersten Satz. [SJC]

A.7 Feature

Das Feature ist die Erweiterung der Nachricht um Analysen und Hintergründe. Der Stil ist ebenfalls eher sachlich, manchmal können aber auch Kommentare einfließen. Der Leser soll umfassend und ausführlich informiert werden, um sich anschließend eine eigene Meinung bilden zu können.

Durch die gezielte Auswahl und Darstellung bestimmter Sachverhalte ist das Feature weniger neutral als die knappe Nachricht.

A.8 Reportage

Die Reportage unterscheidet sich von den anderen journalistischen Texten dadurch, dass sie erzählt statt nur berichtet oder kommentiert. Man könnte sie auch zu den literarischen Gattungen zählen, denn, wie das Handbuch des Journalismus verrät: »Wenn Dichter erzählen, werden ihre Texte zu Kurzgeschichten oder Romanen; wenn Journalisten erzählen, schreiben sie eine Reportage.« [SJC, S.104] Sie schildert subjektiv, anschaulich und detailliert von persönlichen Erlebnissen des Autors und ist chronologisch geordnet. Da sich persönliche Berichte nur im Grad der Professionalität von journalistischen Reportagen unterscheiden, fallen sie ebenfalls in dieses Genre.

B. Literatur

Literarische Texte dienen vor allem der Unterhaltung des Lesers. Im Gegensatz zu den anderen Genres handelt es sich hier um eine *Kunstform*, der Autor möchte der Welt seine Gedanken und Gefühle mitteilen.

B.1 Gedicht

Kennzeichnend für das Gedicht ist, bis auf wenige Ausnahmen, die Versform. Alle Zeilen sind in etwa gleich lang und weisen Regelmäßigkeiten in Silbenzahl und -betonung auf (Versmaß). Oft treten zusätzlich noch Reim, Assonanz (Gleichklang der Vokale) oder Alliterationen auf [WIK]. Die Länge von Gedichten ist sehr unterschiedlich und reicht von wenigen Zeilen bis zu ganzen Büchern. Wegen der ähnlichen Form sind auch Gebete Teil diese Genres, obwohl meist kein festes Versmaß vorliegt.

B.2 Prosa: Roman, Erzählung, Kurzgeschichte

Prosa bezeichnet alle erzählenden Texte, die – im Gegensatz zu Versen oder der typischen Form aus Dialogen und Handlungsanweisungen des Dramas – als Fließtext verfasst werden. Dazu gehören Werke unterschiedlicher Länge wie beispielsweise Romane, Erzählungen oder Kurzgeschichten.

B.3 Drama: Drehbuch, Theater

Das Drama bildet die dritte große Literaturgattung und ist für die Bühnendarstellung gedacht. Oft herrschen Dialoge vor, die um einige Handlungsanweisungen erweitert werden. Während früher hauptsächlich Theaterstücke in diesem Genre zu finden waren, fallen heute auch Hörspiele oder Drehbücher darunter.

B.4 Kurztexte

Neben diesen literarischen Hauptgenres gibt es noch einige sehr kurze unterhaltende Texte wie Rätsel oder Witze, die sich nicht richtig einordnen lassen. Da sie eher selten und so gut wie nie einzeln auftreten, wurden sie nicht berücksichtigt.

C. Information/Wissen

Texte dieser Kategorie haben die Vermittlung von Wissen zum Ziel. Anders als beim Journalismus handelt es sich hier nicht um aktuelle, sondern um zeitlose Informationen.

C.1 wissenschaftlicher Bericht

In einem wissenschaftlichen Bericht werden auf wenigen Seiten neue Forschungsarbeiten beschrieben. Die Leser sind meist aus verwandten Fachbereichen, das Niveau entsprechend hoch.Oft folgen sie einer festen Gliederung aus Abstract (Zusammenfassung), Einleitung, Methoden, Ergebnisse, Diskussion, Summary und Literaturliste.

C.2 Erklärungen

Längere und eher an die Allgemeinheit gerichtete Texte, die bestimmte Dinge oder Sachverhalte erklären oder beschreiben, sind in diesem Genre zusammengefasst. Beispiele sind Schulbücher, Artikel aus Enzyklopädien (z.B. Wikipedia) oder Texte aus Sachbüchern.

C.3 Anleitung

Anleitungen enthalten Anweisungen an den Leser, die er befolgen muss, um ein gewünschtes Ziel zu erreichen. Typische Vertreter sind Kochrezepte, Tutorials oder »Do it yourself«-Tipps.

C.4 FAQ

Ähnlich wie Anleitungen dienen auch FAQ-Seiten häufig der Lösung von Problemen des Lesers. Allerdings sind sie anders aufgebaut: Die Probleme sind als Fragen formuliert, die beantwortet werden. Dadurch ist es möglich, schnell die gewünschte Information zu finden.

C.5 Lexikon

In einem Lexikon findet man kurze Informationen zu einer Reihe oft alphabetisch sortierter Begriffe. Hierunter fallen auch Computersprachreferenzen oder Glossare.

C.6 Zweisprachiges Wörterbuch

Ein Spezialfall von Lexika sind zweisprachige Wörterbücher. Der Unterschied ist, dass zu den einzelnen Wörtern keine Erklärungen, sondern Übersetzungen gegeben werden.

C.7 Präsentation, Werbung

In diese Klasse fallen eher subjektive Darstellungen von Informationen wie Produktbeschreibungen, Unternehmensprofile (»Wir über uns«), persönliche Homepages und Werbung.

C.8 Statisitken

Statisitken sind Dokumente, bei denen die Information hauptäschlich aus Zahlen besteht – oft in Form einer Tabelle, aber auch eingebettet in Texte.

C.9 Code

Dieses Genre enthält Code-Listings.

D. Dokumentation

Hier sind Genres versammelt, die auf verschiedene Arten Anweisungen, Begebenheiten oder Ereignisse von öffentlichem Interesse dokumentieren.

D.1 Gesetze und Regeln

Diese Textart umfasst meist offizielle Regelwerke wie Gesetze, Vereinssatzungen, religiöse Gebote

oder Spielregeln. Sie dokumentieren verbindliche Handlungsanweisungen oder Verbote für bestimmte Personengruppen.

D.2 Offizieller Bericht

Offizielle Berichte werden von oder für Organisationen verfasst und informieren über vergangene Begebenheiten. Sie decken entweder bestimmte Zeiträume ab (z.B. Jahresberichte) oder ein bestimmtes Ereignis (z.B. Sitzungsbericht). Oft ist es auch gesetzlich vorgeschrieben, solche Berichte zu veröffentlichen.

D.3 Protokolle

Protokolle dokumentieren den zeitlichen Ablauf bestimmter (offizieller) Geschehnisse. Es werden kurz wesentliche Handlungen beschrieben, jedoch hauptsächlich die Gesprächsbeiträge der einzelnen Personen aufgezeichnet. Beispiele sind Meeting Minutes und Gerichtsprotokolle.

D.4 Zitate

Sammlungen von Aussprüchen, oft von einer bestimmten Person, aus einem künstlerischem Werk oder zu einem ausgewählten Thema, bilden dieses Genre. Die Zitate können aus wissenschaftlichen oder literarischen Texten stammen oder mündliche Äußerungen dokumentieren.

E. Verzeichnis/Directory

Texte dieser Klasse präsentieren Informationen in Form von Listen. Sie dienen meist als eine Art Nachschlagewerk.

E.1 (juristische) Personen

Verzeichnisse von Organisationen, Personen, Firmen, Ländern, Städten etc. sind hier zusammengefasst. Diese Listen sind oft alphabetisch sortiert.

E.2 Katalog

Ein Katalog ist eine Auflistungen von Waren oder anderen zum Verkauf bestimmter Dinge, die durch die Angabe von Preisen gekennzeichnet ist.

E.3 Ressourcen

Diese Listen enthalten Verweise auf andere Texte, also Linksammlungen und bibliographische Referenzen/Literaturverzeichnisse.

E.4 Timelines

Timelines sind Texte, die eine chronologisch gegliederte Aufzählung von Ereignissen oder Veranstaltungen enthalten. Hierzu gehören zum Beispiel Lebenslauf, Terminkalender oder die Auflistung historischer Begebenheiten.

E.5 Wortliste

Wortlisten sind reine Auflistungen von Begriffen, teilweise für linguistische Zwecke. Meist handelt es sich aber um Spamseiten ohne Inhalt, die nur dazu dienen, von Suchmaschinen gefunden zu werden. Auf solchen Seiten findet man oftmals eine Anhäufung von Schreibfehlern.

E.6 Sonstige Listen

Alle anderen Verzeichnisse sind in diesem Punkt zusammengefasst. Es handelt sich hier im weitesten Sinne um Dinge, beispielsweise Fotogalerien, Projektbeschreibungen oder Stellenanzeigen. Eine weitere Unterteilung wäre möglich, erscheint jedoch nicht sinnvoll, da es sich eher um Unterschiede im Inhalt statt in Form oder Zweck handelt.

F. Kommunikation

In diesem Bereich sind alle Texte zusammengfasst, deren Zweck die *persönliche* Kommunkation ist. Hierin unterscheidet sie sich von den anderen Kategorien, in denen die für Massenmedien typische unidirektionale Vermittlung von Inhalten stattfindet.

F.1 Brief/Mail/Rede

Bei Texten dieses Genres findet eine direkte, unidirektionale Kommunikation zwischen Verfasser und Empfänger statt, der Leser oder Zuhörer wird persönlich angesprochen. Im Falle des Leserbriefs sind die klassischen Rollen vertauscht: hier teilt der Medienkonsument dem Produzenten etwas mit.

F.2 Forum, Gästebuch

Forum und Gästebuch sind Internet-spezifische Genres. Durch spezielle Eingabeformulare können mehrere Menschen neue Beiträge verfassen oder auf andere antworten. Während Gästebücher eher dazu dienen, dem Verfasser der Website Nachrichten zu hinterlassen, finden in Foren Diskussionen oder Gespräche zwischen den Teilnehmern statt. Sie sind häufig nach Themengebieten gegliedert und beschränken sich auf einen bestimmten Bereich (Musik, Programmiersprachen, Selbsthilfegruppen).

F.3 Blog

Blogs (als Kurzform für »Weblogs«) findet man ebenfalls nur im Internet. Ein Autor verfasst ähnlich wie in einem Tagebuch Beiträge, die von den Besuchern der Seite kommentiert werden können. Sehr oft werden persönliche Erlebnisse und Beobachtungen geschildert, man findet aber auch sachliche Blogs, die sich mit einem bestimmten Thema beschäftigen und Neuigkeiten hierzu veröffentlichen.

F.4 Formulare

Über Formulare kann der Betrachter einer Seite mit deren Autoren kommunizieren. Typisch hierfür sind Bestellungen, Anmeldungen, oder Kontaktaufnahme. Auch an die Seite selbst können Anfragen gestellt werden, wie zum Beispiel bei der Suche.

G. Nichts

Unter »Nichts« sind alle Seiten ohne Funktion zusammengefasst: Splash-Pages (Seiten die der eigentlichen Webseite vorgeschaltet werden, oft mit einem Bild o.ä.), leere Framesets, Weiterleitungsseiten, leere Seiten und Fehlermeldungen.

H. Kombination

Hier finden sich Zusammenfassungen mehrerer Texte zu einem Ganzen. Davon ausgenommen sind die Spezialfälle Blog und Forum.

Abbildung 2.1 zeigt, welche Genres üblicherweise in welchen enthalten sein können. Es ist nicht ausgeschlossen, dass in einigen seltenen Fällen noch andere Abhängigkeiten vorkommen. Man sieht, dass einige Klassen als *Container* dienen, das heißt sie stellen einen bloßen Rahmen für die anderen Genres dar. Das Vorhandensein von Teilgenres ist für diese Texte obligatorisch. Eine weitere Gruppe sind alleinstehende Genres, die *optional* auch andere Textarten enthalten können (z.B. wissenschaftliche Texte mit Statistiken, Anleitungen mit Code-Listings). Daneben gibt es *atomare* Klassen, die nicht als Behälter für andere dienen können. Einige davon sind zusätzlich *abgeschlossen*, was bedeutet, dass sie weder Genres enthalten können, noch ein Teil anderer Klassen sind (z.B. Gesetze, Nichts).

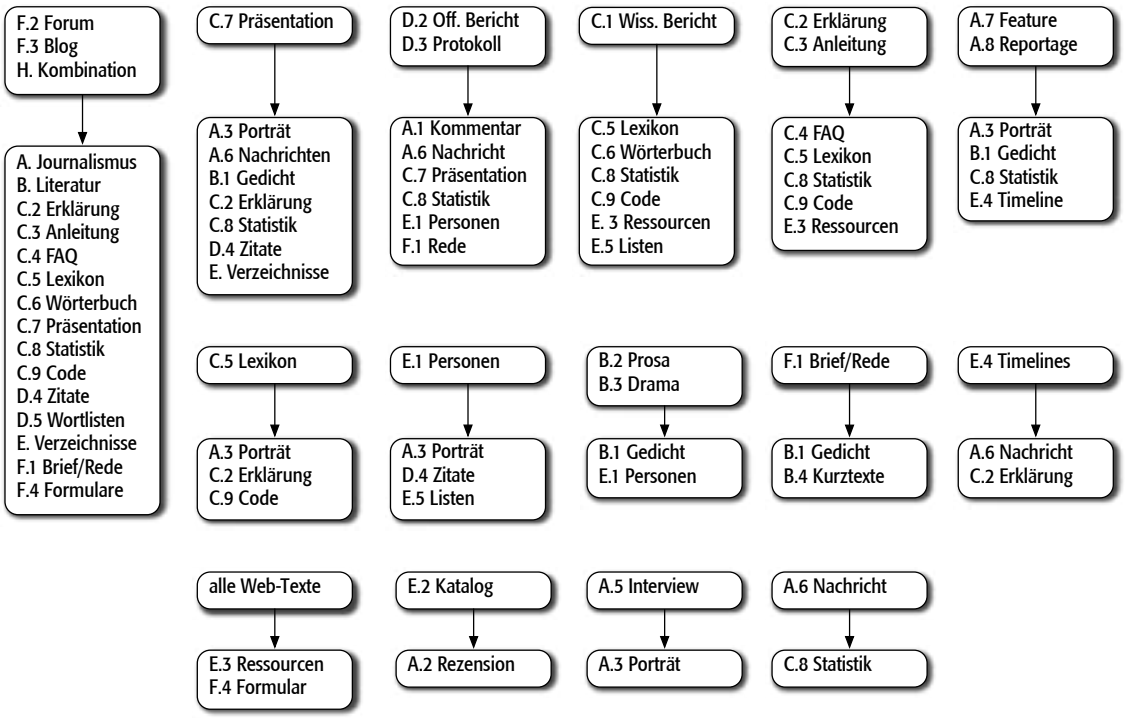

Abbildung 2.1: Meronymie der Genres

Einige der zufällig ausgewählten Texte ließen sich in die Hierarchie nur schlecht einordnen oder landeten in Klassen, wo man sie ursprünglich nicht erwartet hätte. Oftmals liegt das an einem Konflikt zwischen Form und Funktion. So haben beispielsweise nach Datum geordnete historische Schilderungen die Form einer Timeline, aber die Funktion einer Erklärung. Weil eine Mehrfachklassifikation von Texten erlaubt ist, sind solche Grenzfälle aber unproblematisch. Gleiches gilt für Mischformen wie Briefromane. Ein weiterer Grund für Schwierigkeiten bei der Klassifikation besteht darin, dass man intuitiv auch andere Merkmale als Form und Funktion berücksichtigt. So sind sich einige Tagebücher und Autobiografien sehr ähnlich, da beide chronologisch von Ereignissen aus dem Leben des Autors berichten. Dadurch fallen auch die eher unprofessionellen und privaten Tagebücher in den Bereich Literatur, obwohl man dies auf Grund des unterschiedlichen Veröffentlichungs-Rahmens erst nicht für richtig hält.

Manche Texte konnten in keines der Genres eingeordnet werden, darunter Spendenaufrufe oder politische Manifeste. Dies zeigt, dass entweder die Hierarchie nicht lückenlos ist, oder dass die De-

finitionen der einzelnen Genres verbessert und angepasst werden müssen, so dass sie auch die bisher nicht klassifizierbaren Texte umfassen.

2.2.3 Evaluation

Um einen Eindruck zu bekommen, wie gut die Einteilung der Genres und ihre Definition ist, habe ich eine Testperson (einen 33 Jahre alten Grafikdesigner der mit dem Thema vorher nichts zu tun hatte) gebeten, 70 der Texte im Trainkorpus – zwei aus jedem Genre – in die gegebenen Klassen einzuordnen. Durch nachträgliche Änderungen im Klassensystem wurde leider einer der offiziellen Berichte (D.2) gelöscht, tatsächlich wurden also nur 69 Texte sortiert. Die genaue Aufgabenbeschreibung ist im Anhang beigefügt.

Bei 53 der Texte (76,8%) gab es eine Übereinstimmung zwischen den beiden Klassifikationen. Zwei Texte wurden in andere Klasse eingeordnet, die aber auch richtig ist. Zehn der falschen Texte stammen aus einem ähnlichen Genre oder sind Grenzfälle. Absolut falsch sind nur fünf Texte, also 7,2%. Tabelle 2.7 zeigt die Texte, bei denen es uterschiedliche Auffassungen über ihre Genrezugehörigkeit gab.

ist	erkannt als	Bemerkung	Bewertung
C.9	F.2	Forum mit Code	richtig
F.4	G	Fehlermeldung mit Suchfeld	richtig
C.1	C.3	»Tutorial« in Überschrift	ähnlich
C.7	D.2	Firmeninformation mit Zahlen	ähnlich
C.2	D.1	Handlungsanweisungen	ähnlich
A.7	A.6	verwandte Genres	ähnlich
E.1	E.6	verwandte Genres	ähnlich
A.8	F.3	Tagebuch (ähnlich Blog)	ähnlich
A.2	A.1	verwandte Genres	ähnlich
F.1	D.2	Rede im Parlament (Teil eines Sitzungsberichts)	ähnlich
E.6	A.8	Glosse in Listenform	ähnlich
A.1	A.8	erzählt und kommentiert Begebenheit	ähnlich
F.4	G	ein einzelnes Formularfeld	falsch
C.7	G		falsch
A.7	A.8		falsch
A.4	F.1		falsch
E.2	G		falsch

Tabelle 2.7: Nicht in die Original-Klasse eingeordnete Texte

Interessant war, dass einige der Texte (z.B. das einsame Formularfeld) von ersten Versionen meiner Klassifikatoren auf die gleiche Weise falsch eingeordnet wurden. Ansonsten gab es nur wenige Zusammenhänge zwischen den von Hand und den automatisch falsch erkannten Genres. Nur die Klassifikation von Glosse als Brief, sowie Feature und Kommentar als Reportage lässt sich bei beiden beobachten (vgl. Abb. 9.1: Konfusionsmatrix).

Im Vergleich mit bisherigen Genresystemen ist dieses Ergebnis als sehr gut zu bezeichnen, Roussinov et. al. erreichen bei 1076 Texten und 116 Genres eine Übereinstimmung von 49,63% [ROU]. Bei einem Experiment von Crowston und Williams [CRW98] wurden 68% von 790 Texten identisch

eingeordnet und nochmal 10% in ähnliche Klassen. Dabei standen 48 vorher ermittelte Kategorien zur Verfügung, außerdem konnten bei Bedarf zusätzliche Klassen angelegt werden. Das gute Abschneiden meiner Genres hängt auch damit zusammen, dass die Anzahl der Klassen geringer und außerdem fest vorgegeben war. Dadurch gibt es weniger Möglichkeiten, Fehler zu machen.

3 Methode

Neben der Entwicklung der Hierachie liegt der Schwerpunkt dieser Arbeit in der automatischen Erkennung des Genres von Texten. Dafür wurde ein Trainingskorpus angelegt und analysiert, für jede Textart ein Klassifikator geschrieben und schließlich das Ergebnis mit Hilfe von Testdateien überprüft.

3.1 Korpus

Da das Genresystem ganz neu ist, konnte keiner der vorhandenen Korpora verwendet werden. Deswegen wurden für jedes Genre zwei mal 20 (also insgesamt etwa 1400) Dateien gesammelt, mit Ausnahme der Wörterbücher (C.6) und der Wortlisten (E.5), für die nicht genügend gefunden werden konnten. Bei den Wortlisten liegt dies darin begründet, dass die meisten dieser Seiten keine eigentlichen Inhalte aufweisen und deshalb nur schwer mit Suchmaschinen ausfindig gemacht werden können, bei Wörterbüchern an der geringen Anzahl unterschiedlicher Anbieter. In den folgenden Versuchen können Wortlisten daher nicht betrachtet werden, bei den Wörterbüchern habe ich mich auf jeweils zehn Files beschränkt.

Es wird jeweils eine einzelne HTML-Seite als Text betrachtet, nicht ganze Websites oder Framesets. Dadurch kann es passieren, dass nur ein Teil eines Textes vorhanden ist, beispielsweise ein einzelnes Kapitel eines Romans oder die erste Seite eines längeren Interviews. Dem Thema der Magisterarbeit entsprechend wurden nur Texte (oder Verwandtes, wie Formulare) in die Sammlung aufgenommen und keine Websites mit Spielen, Videos etc.

Die Dateien stammen teils aus einem bereits vorhandenen General Corpus (vgl. [STR]), teils wurden sie beim zufälligen Surfen gespeichert oder mit den Suchmaschinen Google und Ask.com gefunden. Es kommen nie mehr als drei Dateien zu einem Genre von einer Website, um nicht deren Eigenheiten zu stark zu gewichten. Andererseits wurde versucht, möglichst Texte in *verschiedenen* Genres aus der selben Quelle zu beziehen, wieder um deren Einfluss auf die Klassifikation gering zu halten. Da ein guter Klassifikator für Textgenres unabhängig vom Inhalt funktionieren soll, wurden darauf geachtet, eine möglichst breite Themenpalette zu wählen [FIN].

Das Web wurde als Grundlage für den Korpus gewählt, da die Texte dort allgemein zugänglich sind. Da theoretisch jeder im Internet veröffentlichen kann, unterliegen die Texte in Inhalt, Stil und Aufbau keinerlei Kontrolle oder Beschränkung. Sie sind also im Gegensatz zu gedruckten Dokumenten ungefiltert und damit vielfältiger und interessanter für die Analyse (vgl. [CRW]). Außerdem sind sie durch die Verwendung von HTML um relativ einfach untersuchbare Informationen zur Struktur erweitert.

Die Sprache ist Englisch, wobei darauf geachtet wurde, Seiten aus verschiedenen Ländern und mit verschiedenen Währungen zu sammeln. Als Codierung wurde UTF-8 gewählt, um alle Sonderzeichen verarbeiten zu können. Die Dateien mussten dazu erst auf ihren Zeichensatz geprüft und anschließend gegebenenfalls umgewandelt werden. Da ich kein Programm finden konnte, welches den Zeichensatz zuverlässig erkannt hätte, verwendete ich ein eigenes Perl-Skript (siehe beigelegte CD). Dieses analysiert den Text auf Byte-Ebene und zählt die möglichen Zeichen in den Codierungen ASCII, ISO-8859-1, ISO-8859-15, Windows Latin 1 (cp1252), Mac OS Roman (cp10000_MacRoman)

und UTF-8. Um ganz sicher zu gehen, wurden Dateien mit einem seltenen Zeichensatz von Hand überprüft.

Dokumente mit sehr viel Werbung oder anderen nicht zum Inhalt gehörenden Elementen wurden bereinigt, da diese die eigentlichen Texteigenschaften verfälschen und damit die Klassifikation erschweren können. Textsammlungen (z.B. mehrere Nachrichten in einem Dokument) wurden in ihre Einelteile zerlegt.

Tagging

Um die Part-of-Speech-Merkmale zu bestimmen, wurde der unter der GNU-Lizenz frei verfügbare TreeTagger vom Institut für maschinelle Sprachverarbeitung der Universität Stuttgart [SCH] verwendet, welcher mit dem Penn Treebank Tagset (siehe z.B. [SAN]) arbeitet.

Alle Dateien wurden mit diesem Programm getaggt, um die POS-Infomationen direkt für die Klassifizierung verfügbar zu machen. Dazu mussten die HTML-Files zunächst vorbereitet werden. Als erstes wurden die Elemente, die nicht als Text dargestellt werden, gelöscht, also alles innerhalb von `<script>`, `<head>`, `<style>`, `<noframes>`, `<select>` oder `<textarea>`-Tags und Kommentaren (`<!-- -->`) sowie alle Tags. Anschließend wurden HTML-Entities wie `&auml;` in UTF-8 Zeichen dekodiert und die Satzzeichen (Frage- und Ausrufezeichen, Doppelpunkt und Strickpunkt) zu Punkten normiert. Zum Schluss wurde der Text in einzelne Wörter – bestehend aus Buchstaben und Apostrophen – Zahlen, Punkte und Kommas zerlegt. Aus dieser Eingabe erzeugt der Tagger eine Liste aus Wort und zugehörigem Treebank-Tag.

3.2 Features und Klassifikatoren

Für jedes Genre mussten die kennzeichnenden Features ermittelt werden. Dafür wurden jeweils die 20 Trainingstexte analysiert und Auffälligkeiten in Struktur, Vokabular, Zeichensetzung etc. als vorläufige Merkmale festgestellt. Diese wurden anschließend für den kompletten Trainingskorpus berechnet, um die Thesen zu überprüfen. Ausgehend von diesen Daten wurde mit Perl für jedes Genre ein eigener Klassifikator geschrieben, für den von Hand Schwellenwerte und Kombinationen der Features festgelegt wurden. In einem iterativen Verfahren wurden so lange neue Features gesucht und das Programm optimiert, bis die Werte für Recall und Precision beide etwa 90% (bzw. 90% und 60% für Fließtexte) auf den Trainingsdaten erreichten. Teilweise konnten bereits gefundene Klassifikatoren verwendet werden. So weist beispielsweise das Vorhandensein von Literaturlisten auf wissenschaftliche Texte hin. Abbildung 3.1 zeigt einen Ausschnitt aus dem so entstandenen Klassifikator für Reportagen.

Manche Genres bestehen aus Texten mit unterschiedlicher Struktur: Ressourcen sind zum Beispiel sowohl Literatur- als auch Linklisten. Hier wurden zwei getrennte Erkenner geschrieben und deren Ergebnisse zusammengefasst. Andere Texte wiederum können in unterschiedlichen Stilen verfasst werden. Kommentare wägen beispielsweise entweder Pro- und Contra-Argumente ab oder stellen sich auf eine Seite, meist in Pamphleten. Diese bösartigen Kommentare sind öfter persönlich geschrieben und verwenden deswegen mehr Pronomen in der 1. Person, ihre Sprache ist lockerer und macht mehr Gebrauch von beleidigenden Ausdrücken. Argumentierende Texte sind dagegen unpersönlich und enthalten etwa gleich viele positive wie negative Adjektive. In solchen Fällen muss der Klassifikator beide dieser Stile berücksichtigen und Texte, die entweder die einen oder

die anderen Bedingungen erfüllen, erkennen. Besonders bei klassischen Journalismus-Genres wird dieser Aspekt deutlich: Reportagen berichten von Menschen oder Reisen, Features werden je nach Textlänge in unterschiedlichen Erzählformen verfasst (kurze Texte im Präsens und nicht aus der Ich-Perspektive, längere in der Vergangeheit). Oft werden bestimmte Stil-Vorgaben auch durch das Vorhandensein anderer Merkmale gelockert, so dürfen Nachrichten nur dann informelle und persönliche Sprachelemente (Pronomen in der 1. und 2. Person, Kontraktionen wie »won't«) verwenden, wenn sie Zitate in direkter Rede enthalten; in Erklärungen dürfen lediglich dann viele Namen vorkommen, wenn sie von geschichtlichen Ereignissen berichten – gekennzeichnet durch Vergangeheitsbezeichner wie »century« oder Jahreszahlen vor 1970.

```
# Textlänge u.ä.:
  $length > 2500 && $length < 45000 && $formular < 10 &&
# ist Text:
  $verb > 18 && $konjunktion > 2 &&
# gegen zu sachlich oder literarisch:
  $adj > 17 && $adjPosNeg > 0.5 && $adjPosNeg < 4 && $kontraktion < 2.5 &&
# gegen Kommentare, FAQ, Interview:
  $argumentativ < 1.3 && $verallgemeinernd < 3.8 && $question < 3 &&
# gegen wissenschaftl., Zeitangaben, zu lockere Sprache:
  $wiss_bigramme < 0.01 && $datum < 0.6 && $informell < 3 &&
# gegen Porträt, geschichtliche Erklärung:
  ($portraitWords < 1 || $name + $he < 7) && $vergangeheitsbez < 1 &&
# eher erste Person, nicht zu viele Namen:
  $we+$i > 1.6 && $he < 8 && $name > 0.5 && $name < 6.5 &&
# entweder eindeutig Past oder Present:
  ($past > $present && $past > 0.2 || $present > 0.2) &&
# über Menschen und Lebewesen...
  ($he > 3 || $name > 4 || $the_lebewesen > 2 ||
# ...oder vergangene Reisen und Abenteuer:
  || ($land > 0.5 && $past > 0.4))
```

Abbildung 3.1: Bedingungen für die Klassifikation als Reportage (aus: find_reportage.pl)

Bei Genres, bei denen auf diese Weise keine identifizierenden Merkmalskombinationen gefunden werden konnten, wurden automatisch erstellte Entscheidungsbäume zur Hilfe genommen. Zu deren Erstellung wurde Weka, einer Reihe von Programmen unter GNU-Lizenz für maschinelles Lernen von der University of Waikato [WEKA], verwendet. Dazu wurden aus dem Trainingskorpus zwei Klassen erstellt, eine enthält das gerade bearbeitete Genre, die zweite alle anderen als solches erkannten Texte. Für diese Klassen wurden einige Merkmale extrahiert und daraus der Entscheidungsbaum berechnet. Besonders gut trennende Features wurden testweise in den Gerne-Erkenner übernommen und, falls das Ergebnis zufriedenstellend war, beibehalten. Auf diese Weise sind auch Features in den Klassifikatoren enthalten, die sich nicht theoretisch begründen lassen. Konkret ist dies aber nur der Fall bei Rezensionen, wo entweder viele Adjektive oder wenige Pronomen in der 1. Person verlangt werden, und bei Erklärungen (viele definite Artikel oder wenige Namen). Es wird erwartet, dass diese Klassifikatoren eher schlecht abschneiden.

Für das Genre »Kombinationen« ist die Methode der iterativen Merkmalsbestimmung nicht möglich, weil diese eigentlich kein eigenes Genre sind, sondern vielmehr eine Zusammenstellung mehrerer anderer. Hier wäre eine andere Herangehensweise nötig: Der Text müsste in einzelne Abschnitte zerlegt und deren Textart mit Hilfe der anderen Klassifikatoren bestimmt werden. Da dies jedoch eine komplett neue und auch komplexe Aufgabenstellung wäre, wird die Erkennung von Kombinationen in dieser Arbeit ausgeklammert.

Für weitere drei Genres gibt es ebenfalls keine Klassifikatoren. Zum einen für Wortlisten, für die wie oben erwähnt nicht genug Texte gefunden wurden, zum anderen für »Zitate« und »Sonstige Listen«, da sich diese als recht schwierig erwiesen und in der gegebenen Zeit nicht mehr fertiggestellt werden konnten.

Wortlisten

Häufig werden Wortlisten zur Bestimmung der Features benötigt. Einige waren bereits vorhanden und konnten, gegebenenfalls nach einer Säuberung, direkt verwendet werden. Die meisten mussten jedoch durch eigene Recherchen zusammengestellt werden. Verwendet wurde dazu eine elektronische Version des Oxford American Dictionary, verschiedene Quellen im Internet und Analysen mehrerer Texte in Zeitungen, dem Trainkorpus etc. Mit Ausnahme der natürlich beschränkten Listen (z.B. Pronomen) sind die Listen unvollständig.

Eine Art von Wortlisten umfasst Subkategorien von Part-of-Speech. Dazu gehören sehr kleine und begrenzte Listen wie Personalpronomen, nach Genus, Numerus und Person aufgeteilt, und definite bzw. indefinite Artikel sowie Konditional-Verben. Die anderen Sammlungen waren erheblich aufwändiger zu bestimmen. So wurden Listen von positiven, negativen und neutralen Adjektiven erstellt, indem aus dem Oxford-Dictionary sämtliche Adjektive extrahiert und von Hand kategorisiert wurden. Um die Arbeit etwas einzuschränken, wurden davor all jene entfernt, die nicht zu den 200 000 am häufigsten verwendeten Wörtern des Englischen gehören. Für Synonyme für »sprechen« und Verben der Wahrnehmung wurden mit Hilfe eines Thesaurus (ebenfalls von Oxford American Dictionaries) gesammelt.

Eine weitere Kategorie betrifft die Art der verwendeten Sprache. Hierzu gehören eine schon vorhandene Aufzählung der 200 000 häufigsten Wörter sowie einige unvollständige Listen, darunter eine mit lockeren/informellen Beriffen (z.B. »folks«, »well«), eine mit zum Argumentieren (»finally«, »nevertheless«) oder Folgern (»because«, »hence«) verwendeter Sprache, sowie Sammlung von beleidigenden (»fool«, stupid«, »crazy«), vagen (»something«), verallgemeinernden (»they«, »always«), zustimmenden/ablehnenden und deiktischen Wörtern. Diese entstanden durch die Auswertung einiger Texte und wurden durch Zusatzinformationen, zum Beispiel aus Hinweisen zum Verfassen von Erörterungen im Englischen, erweitert. Aus dem Dictionary wurden alle altertümlichen Wörter, zu erkennen an dem Zusatz »archaic« in der Begriffserklärung, extrahiert.

Daneben gibt es noch Listen von Personen-, Städte- und Ländernamen, solche mit typischen Firmenbezeichnungen oder Anreden/Titeln (»Mr«, »Professor«), eine Sammlung von Akronymen und Emoticons und – da in literarischen Werken auch oft »der Räuber«, »die Mutter« oder »das Kaninchen« als Personen vorkommen – eine Liste von Berufen, Tieren und Verwandschaftsbezeichnungen. Als Quelle diente hier hauptsächlich Wikipedia, wo es zu vielen Themen Listen von Begriffen gibt. Die Vornamen stammen aus einer bereinigten Liste vom CIS an der LMU München. Bei der Entscheidung, ob ein Wort in die Liste aufgenommen werden sollte oder nicht, stand im Vordergrund, ob es gleichzeitig auch ein häufig im Englischen verwendeter Begriff ist. Wären diese Teil der Listen, so würden viele normale Wörter als Namen (z.B. »Beat«) oder Akronyme (»BAD«) erkannt und dadurch das Ergebnis verfälscht.

Bestimmte für ein Gerne spezifische Schlüsselwörter wurden durch eine oberflächliche Analyse der Texte im Trainkorpus identifziert. Dazu gehören zum Beispiel Fehlermeldungen (»File not Found«

oder »Seeing this instead of the website you expected«), die Benennung von Artikeln und Absätzen in Gesetzestexten (»Article 3«) oder Überschriften (»Abstract«, References«) und Bigramme (»our results«, »we propose«) in wissenschaftlichen Texten. Eine besonders umfangreiche Liste ist mit 119 Einträgen die der Wörter in Codelistings. Dafür wurden Referenzen und Codelistings in den Programmier- und Skriptsprachen Java, Delphi, Python, Actionscript, Javascript, Perl, C, C++, C# und PHP untersucht und die Schlüsselwörter gesammelt. Wörter, die gleichzeitig auch Teil der englischen Sprache sind (»while«, »private«), wurden anschließend wieder entfernt.

Alle längeren Wortlisten befinden sich auf der beigefügten CD, die kürzeren, insbesondere die Keywords, stehen in Kapitel 4 bei der Feature-Auflistung.

Automatische Verfahren

Zusätzlich zu den einzelnen Klassifikatoren der Genres und deren Kombination (vgl. 5) wird noch untersucht, wie gut die gefundenen Merkmale sich zur automatischen Gerne-Erkennung eignen. Dazu werden mit einem einzigen Programm sämtliche Features berechnet und zeilenweise je Text ausgegeben. Aus diesen Daten bestimmen die Klassifizierungs-Algorithmen ein Model, welches anschließend mit den Testdateien überprüft wird. Es werden Naive Bayes, K-Nearest-Neighbour, Entscheidungsbäume und Support-Vector-Machines untersucht. Die Funktionsweise dieser Verfahren wird in Kapitel 5 beschrieben.

4 Features

In diesem Abschnitt wird zunächst dargestellt, welche Arten von Features in bisherigen Arbeiten verwendet und wie diese bestimmt werden. Es folgt eine Einteilung der Merkmale in verschiedene Kategorien. Anschließend werden die in dieser Arbeit benutzten Features für jedes einzelne Genre aufgelistet sowie Probleme und Lösungsansätze für deren Extraktion aus den Texten geschildert.

4.1 Bisherige Arbeiten

In der Literatur zum Thema findet man große Unterschiede, welche Merkmale zur Klassifikation herangezogen und mit welcher Methode sie gefunden werden. Teils entstehen sie durch statistische Analysen von Korpora und sind damit eher deskriptiv [BIB, ILL], teils werden sie linguistisch begründet und stellen die Erwartungen der Verfasser bezüglich der Verwendung in bestimmten Textarten dar, sind also präskriptiv [WHI].

Einer der ersten, die stilistische Unterschiede von Texten untersuchte, war Biber [BIB]. Er analysierte die Häufgkeitsverteilung von 67 *Part-of-Speech*-Merkmalen in knapp 5 000 englischen Texten und generierte mit Hilfe der Faktor-Analyse Paare von komplementären Feature-Gruppen, die jeweils eine Dimension bilden. Kommen Features aus einer der beiden Gruppen häufig vor, so sind die der anderen sehr selten vertreten. Im entstandenen mehrdimensionalen Feature-Raum können mit Clusterverfahren automatisch Textkategorien gebildet werden, die allerdings keine Entsprechung in Genres finden.

Ein ähnliches Vorgehen wird im TyPTex-Projekt [ILL] verfolgt, allerdings wird hier in einem gegebenen System untersucht, welche Features in welchen Klassen eher häufig oder seltener vorkommen. Dabei handelte es sich zwar um sechs thematische Gebiete aus der französischen Tageszeitung »Le Monde«, aber da kein spezielles Vokabular untersucht wird, müsste diese Methode auch für Genres funktionieren.

In vielen Arbeiten (u.a. [KAC, DEWE]) werden auch *statistische* Merkmale wie Textlänge, Anzahl der Sätze, durchschnittliche Wort- und Satzlänge oder Type-Token-Ratio betrachtet. [KES] und [DEW] berücksichtigen zusätzlich noch Varianz und zusammengesetzte Features wie Satzkomplexität (berechnet aus Silbenzahl, Wortlänge und Satzlänge).

Eine weitere Möglichkeit ist die Betrachtung von speziellem *Vokabular* oder ähnlichem: [DEWE] sucht beispielsweise betonende und abschwächende Ausdrücke, [ROU] Telefonnummern und Preise, [REH] setzt Schrift- und Sprechsprache ins Verhältnis, zählt Formularfelder und untersucht Bildinhalte, [KES] betrachtet Satzzeichen, Akronyme und Wörter mit lateinischen Affixen und [DEW] die Namen von Sternzeichen und Wochentagen. Neben Part-of-Speech ist dies die am häufigsten verwendete Art von Features.

Die meisten unterschiedlichen Merkmale betrachtet [DEW]. Neben den oben bereits genannten sind dies Zeitänderungen im Textverlauf und die *Formatierung* von Texten, zum Beispiel Einrückungen und Zeilenabstand.

Im Gegensatz dazu legen Kessler et. al. Wert auf eine einfache Erkennbarkeit der Features und beschränken sich deswegen auf lexikalische Merkmale wie Interpunktion und bestimmte Wortgrup-

pen wie zum Beispiel Anreden, Akronyme oder solche mit lateinischen Affixen sowie zahlreiche Kombinationen [KES].

Noch einfacher machen es sich Stamatatos et. al. und wählen als Features die 30 häufigsten Wörter im BNC (British National Corpus) sowie 8 Satzzeichen. Ihr Ziel dabei ist es, mit möglichst schlichten Mitteln und einem kleinen Trainkorpus (für jedes ihrer 4 Genres 20 Dateien) gute Ergebnisse zu erzielen. [STA]

[WAS] untersuchen für schwedische Texte, ob Wort- oder Part-of-Speech-Features (mit und ohne Subkategorien) besser für die Genre-Klassifikation geeignet sind. Dazu bestimmen sie deren Häufigkeiten in Uni-, Bi- und Trigrammen. Das beste Ergebnis bei einer Naive-Bayes-Klassifikation erzielen die Subkategorie-POS-Merkmale in Bigrammen.

[ROU] konzentrieren sich auf Struktur- und HTML-spezifische Merkmale. Dazu gehören die typische Frage-Antwort-Abfolge von FAQs und die Gliederung von wissenschaftlichen Texten genauso wie allgemeine Aussagen über den hierarchischen Aufbau des Textes. Auch Kontextinformationen wie Metadaten, die Anzahl eingehender Links und die URL (enthält sie Zahlen? besteht sie nur aus dem Hostnamen?) werden betrachtet.

Mehrere Arbeiten [ROU, SHE, CRW] werten die Anzahl von Formularelementen, Links und Bildern aus. Rehm will sogar Bildinhalte analysieren und wählt auch sonst schwer bestimmbare oder eher aussagelose Features wie den HTTP-Header, die DTD des Dokuments, Javascript oder die Position in der Datei innerhalb der kompletten Website. Ein interessantes, aber ebenfalls nur mit hohem Aufwand erkannbares Merkmal ist die Anzahl der Rechtschreibfehler [REH], die in von Laien verfassten Texten vermutlich höher ist als in professionellen.

Deixis, Zeit und Involviertheit

De Saint-Georges [ISG] untersuchte den Gebrauch von Deixis und verschiedenen Zeitformen für persönliche Homepages. Deixis nimmt Bezug auf den Kontext, in welchem Äußerungen stattfinden. Sie tritt in den drei Dimensionen Person, Zeit und Ort auf – wer spricht, wann und wo? – und äußert sich im Gebrauch von Pronomen der 1. und 2. Person, temporalen Adverbien (jetzt, morgen, heute, gestern) und komplexeren Ausdrücken, die sich auf den aktuellen Zeitpunkt beziehen (letzten Monat, dieses Jahr) sowie relativen Ortsangaben (hier, dort, dieses).

Treten viele deiktische Wörter in einem Dokument auf, so deutet das auf einen hohen Grad an *Involviertheit* hin, also das Einbeziehen des Lesers [ARG]. Textarten, die diese Interaktion wünschen, sind alle meinungszentrierten journalistischen Texte sowie die Kommunkationsgenres und literarische Werke. Sachliche Texte setzen solche Elemente eher sparsam ein. Argamon et al. zählten die Vorkommen von Pronomen in der 1. bzw. 2 Person und fanden in Fiction-Texten circa 290 bzw. 240 pro 10 000 Tokens, in Non-Fiction-Texten nur knapp 120 bzw. 75. Auch für andere »Involvedness-Features«, zum Beispiel Negationen (ca. 60/110), oder Kontraktionen (ca. 20/90) kann man starke Unterschiede erkennen. [ARG]

Auch bei der Verwendung von Zeiten unterscheiden sich fiktionale von sachlichen Genres, allerdings nicht so stark. Das Unmittelbarkeit ausdrückende [ISG] Präsens findet man in sachlichen Texten etwas seltener, nämlich ca. 280 mal im Vergleich zu knapp 320. [ARG]. Eine Analyse der Vorkommen von Orts- und Zeitdeixis in den Trainingstexten ergab, dass ein Zusammenhang zu der von

Argamon untersuchten Personen-Deixis besteht: Auch hier weisen kommunikative, überzeugende oder literarische Texte eine höhere Zahl dieser Wörter auf (vgl. Tabelle 4.1). Diese Erkenntnisse weisen darauf hin, dass diese Art von Merkmalen bei der Genre-Erkennung hilfreich sein kann.

Genre	Ort	Zeit	Involv.	Genre	Ort	Zeit	Involv.	Genre	Ort	Zeit	Involv.
A.1	6,579	11,366	+	C.2	7,379	4,209	0	E.1	3,142	2,334	-
A.2	7,357	2,902	0	C.3	10,034	3,272	-	E.2	4,559	2,778	-
A.3	6,336	3,714	0	C.4	8,221	4,020	0	E.3	2,965	1,871	-
A.4	8,921	9,136	+	C.5	5,962	3,041	-	E.4	6,101	3,853	-
A.5	13,546	8,084	+	C.6	1,219	7,259	-	E.6	6,102	4,745	-
A.6	2,804	2,421	-	C.7	4,102	2,961	0	F.1	8,409	7,447	+
A.7	6,391	4,496	0	C.8	2,275	0,947	-	F.2	6,731	7,954	+
A.8	11,498	7,137	+	C.9	2,092	3,988	-	F.3	8,449	7,318	+
B.1	10,271	8,061	+	D.1	8,431	0,839	-	F.4	4,267	8,068	-
B.2	10,686	7,693	+	D.2	4,633	1,937	-	G	5,963	3,014	-
B.3	14,123	13,214	+	D.3	4,477	1,658	-				
C.1	6,016	2,519	-	D.4	8,860	8,326	0				

Tabelle 4.1: Mittelwerte der Vorkommen (in 10 000) von Orts- (here, there) und Zeitdeixis (now, today, yesterday, tomorrow) relativ zur Textlänge und ungefährer Wert der Involviertheit (+ stark, o neutral, - schwach)

Whitelaw und Argamon [WHI] versuchen, die Auswahl ihrer Features theoretisch zu untermauern und untersuchten unter anderem die sogenannte *interpersonelle Metafunktion*, die in etwa der Involviertheit entspricht und Pronomen, einleitende Floskeln wie »I suppose«, »in general« oder »unfortunately« und Modalität umfasst. Letztere lässt sich wiederum aufteilen in Konjunktiv, einschränkende (»perhaps«) oder verallgemeinernde (»always«) Wörter und ähnliches. Es wird angenommen, dass der Autor aus den möglichen Formulierungen mit gleicher denotationeller Bedeutung genau die auswählt, die die von ihm gewünschte Funktion am besten erfüllen. Damit folgen sie den Prinzipien der *Systemic Functional Linguistics*, die besonderes Augenmerk auf die konnotative Bedeutung von Texten richtet.

Einen ähnlichen Ansatz verfolgen auch [DEWE] und betrachten lexikalische, syntaktische und strukturelle Merkmale als Wahl des Autors. Als Grundlage dienen allerdings dennoch die rein deskriptiven Features von Biber [BIB].

4.2 Bestimmung der relevanten Features

Inspiriert durch die Ergebnisse bisheriger Arbeiten wurde für jedes Genre eine Reihe relevanter Merkmale bestimmt. Dazu wurden, wie in Kapitel 3 beschrieben, die Texte von Hand analysiert und zuvor intuitiv aufgestellte Thesen zu Sprachgebrauch und Form der Textarten überprüft. Oft waren die Annahmen zu bestimmten Texten richtig (und teilweise trivial): Kataloge enthalten tatsächlich viele Preisangaben und Briefe eine Anrede und Verabschiedung. In Porträts werden überwiegend positive Adjektive und Personalpronomen eines bestimmten Geschlechts verwendet, die Sprache der meinungsorientierten journalistischen Texte Glosse und Kommentar, sowie in Grenzen Interview und Feature, ist lockerer als die in sachlich informierenden Texten. Interviews und Drehbücher, die beide eine ähnliche Form aufweisen, unterscheiden sich in der Anzahl der Gesprächsteilnehmer; Konjunktionen dienen als Indikator für Fließtext-Genres (im Gegensatz zu Listen u.ä.). Auf die Verwendung von unbegründbaren statistisch ermittelten Merkmalen wurde, abgesehen von

den beiden oben (3.2) genannten Ausnahmen, verzichtet; vor allem weil der Trainingskorpus zu wenig umfangreich ist, um verlässliche Vorhersagen treffen zu können.

Wenn es zu vielen Fehlklassifikationen von Texten aus bestimmten anderen Genres kam, wurden teilweise *Negativ-Merkmale* eingeführt. Um Anleitungen zu filtern, können zum Beispiel alle Dokumente, die eine bestimmte Anzahl an Maßangaben überschreiten, ausgeschlossen werden.

Manche Features sind *Kombinationen* anderer Merkmale. Um zu erkennen, ob ein Text argumentierend ist, dienen beispielsweise die Anzahl der Fragen, Negationen, argumentierenden, verallgemeinernden, kausalen und konditionalen Wörter als Kennzeichen, aus welchen ein Gesamtwert bestimmt wird. Teilweise können Merkmale auch ein *indirekter* Indikator sein: Direkte Rede kann durch die Anzahl der Anführungszeichen gemessen werden, informelle Sprache durch die Verwendung von Kontraktionen oder die Involviertheit durch die oben bereits genannten Personalpronomen und deiktischen Begriffe.

Auffällig war, dass einige Genres schon allein durch ihr spezifisches Vokabular erkannt werden konnten, andere hingegen, wie beispielsweise die verschiedenen Listen, benötigen eine aufwändige Suche nach Strukturmustern oder, wie die journalistischen und literarischen Texte, eine Analyse von Part-of-Speech-Features. Der Zeitaufwand für das Finden der einzelnen Klassifikatoren und leider auch deren Qualität schwankt daher erheblich.

Interessant ist außerdem, dass die Genres in meiner Einteilung zwar zu großen Teilen durch ihre *Funktion* bestimmt werden, es jedoch die *Form* ist, welche man automatisch erkennen kann. Hieran wird deutlich, dass diese beiden Merkmale von Genres nicht unabhängig voneinander sind; um einen bestimmten Zweck zu erfüllen, wird eine bestimmte Struktur und/oder Schreibstil gewählt. Bei einigen Textarten, zum Beispiel Verzeichnissen, Gedicht, Forum oder FAQ steht allerdings der Formaspekt im Vordergrund. Wie man später bei der Evaluation sehen kann, führt diese Eigenschaft zu besseren Ergebnissen bei der Klassifikation.

4.2.1 Arten von Features

Oberflächenmerkmale

Features in dieser Kategorie abstrahieren von den Inhalten und betrachten lediglich das Aussehen des Textes. Dazu gehören die Struktur, Formatierungen und HTML-Zusatzinformationen. Typische *Strukturmerkmale* sind statistische Informationen wie Text- und Zeilenlänge, Vorkommen von Listen, Bildern oder Tabellen oder die Gliederung in Absätze und ein hierarchischer Aufbau mit Überschriften.

Formatierungen legen fest, wie die Texte dargstellt werden. Hierzu gehören unter anderem Schriftart und -größe, Auszeichnungen (fett, kursiv, in Großbuchstaben, Farbe, Unterstreichungen, Rahmen etc.) und Ausrichtung (zentriert, linksbündig). Speziell für HTML-Seiten gibt es außerdem noch die Angabe, Leerzeichen und Zeilenumbrüche im Quelltext zu interpretieren (`<pre>` und ähnliche).

Schließlich können aus einem HTML-Dokument auch noch nicht dargestellte *Zusatzinformationen* gewonnen werden. Das beste Beispiel dafür sind die sogenannten Meta-Tags, die einen Text um Keywords und Angaben zu Autor und Erstellungsdatum anreichern können. Da die Verwendung dieser Tags jedoch nicht einheitlich gehandhabt wird und (zumindest früher) oft beliebige Keywords erfunden werden, um in Suchmaschinen eine gute Platzierung zu erreichen, werden diese

hier nicht berücksichtigt. Auch der im `<title>` angegebene Seitentitel wird nicht benutzt, da einige der Dateien im Korpus anhand von Stichwörtern gesucht wurden, die im Titel auftreten. Dadurch ist diese eigentlich gute Information nicht mehr verwendbar. Weitere Features sind die Angabe von RSS-Feed-Quellen (»Rich Site Summary«, oder in neuen Versionen »Really Simple Syndication«) und die *Content-to-Code-Ratio (CCR)*, die angibt welcher Abteil des Quellcodes tatsächlich als Text dargestellt wird.

Einfluss der Form

Toms und Campbell [TOC] haben in einem Versuch gezeigt, dass der Einfluss von Struktur- und Darstellungs-Merkmalen bei der Erkennung von Genres durch Menschen nicht unerheblich ist. Dazu entfernten sie aus einer Reihe von Texten einmal den Inhalt, indem sie alle Buchstaben und Zahlen durch X bzw. x und 9 ersetzten, und einmal sämtliche Formatierungen und Struktur. Die Texte ohne Inhalt wurden immerhin zu 56% richtig erkannt, die ohne Form mit 70%. Wenn Inhalte in eine andere, sehr spezifische Struktur gebracht wurden (z.B. ein Brief in Gestalt eines Wörterbuchs), so ist der Einfluss der Form auf die Einordnungs-Emtscheidung so groß, dass die Texte entsprechend dieser Form klassifiziert werden. Deswegen sollten Features dieser Art immer dann berücksichtigt werden, wenn die Struktur eines Genres besonders charakteristisch ist.

Vokabular, POS und Patterns

Einige Genres zeichen sich durch ein spezifisches Vokabular oder einen besonderen Sprachstil aus. Bestimmte Begriffe tauchen unabhängig vom Inhalt des Dokuments in manchen Genres häufiger auf als in anderen. Ein triviales Beispiel hierfür sind Personenlisten, in denen sehr viele Namen und Titel vorkommen. Auch das Nennen des Genre-Namens ist oft ein gutes Kennzeichen (z.B. bei Glossaren oder FAQ-Seiten). Allerdings habe ich versucht dieses Merkmal, aus den gleichen Gründen wie den HTML-Dokumenten-Titel, nicht zu verwenden. Auf Seiten, deren Form relativ fest vorgegeben ist, findet man ebenfalls häufig bestimmte Keywords wie »posted« in Blogs oder »topic« in Foren. Ein anderes Beispiel sind Code-Listings. Hier ist es theoretisch möglich, eine Liste aller Funktionsnamen und ähnliches der wichtigsten Programmiersprachen zu erstellen. Problematisch ist, dass viele dieser Namen auch im normalen Sprachgebrauch sehr häufig verwendete Wörter sind. Deshalb werden diese, wie in Kapitel 3 erwähnt, weggelassen. Daraus ergibt sich zwangsläufig, dass in einigen Code-Listings weniger oder überhaupt keine Schlüsselwörter mehr erkannt werden. Eine Lösung wäre hier, auch den Kontext dieser Begriffe zu betrachten, um zwischen Verwendung als englisches Wort und Teil der Computersprache zu disambiguieren.

Nicht nur einzelne Wörter, sondern auch Bigramme oder längere Wortketten können als Merkmale dienen. Beispiele hierfür sind unbenannte Personen in literarischen Texten, die oft mit »the« gefolgt von einem Lebewesen wie »mouse« oder »blacksmith« bezeichnet werden, feste Begriffe wie »Frequently Asked Questions« oder bestimmte Wendungen wie »I don‘t think so« oder die unpersönliche Briefanrede »To whom it may concern«.

Neben solchen Begriffs-Vorkommen kann man auch noch Häufungen von Wörtern eines bestimmten *Vokabulars* feststellen, beispielsweise altertümliche Begriffe in Romanen oder Fremdwörter (bzw. Wörter die nicht im General English vorkommen) als wichtigen Hinweis für zweisprachige Wörterbücher.

Weniger offensichtlich ist die Verwendung oder das Fehlen bestimmter *Wortarten* wie Pronomen einer bestimmten Person. In Texten, die direkte Rede verwenden, und Reportagen sieht man sehr häufig Pronomen in der 1. Person singular, wissenschaftliche Texte machen regen Gebrauch von der 1. Person plural. Wertende Adjektive trifft man oft in Romanen oder Biografien an, wobei besonders in Nachrufen die positiven Eigenschaften stark überwiegen (bei 75% der Texte im Trainkorpus). Dies lässt sich leicht dadurch erklären, dass man »nicht schlecht über Tote spricht«.

Einige Merkmale berücksichtigen nicht die Wörter an sich, sondern die Art und Weise wie diese aufgebaut sind. Dazu gehören die typischen Variablennamen beim Programmieren, die oft in CamelCase (Binnengroßschreibung), mit Unterstrichen oder – vorgegeben durch die Sprachsyntax – mit vorangestelltem Dollarzeichen geschrieben werden (myVariable, my_variable, $variable). Ein anderes Beispiel sind Zeichenhäufungen, wie man sie oft in unprofessionellen und informellen Texten findet (»Haaaallo«, »Danke!!!!«). Außerdem gibt es auch noch Elemente, die keine Wörter sind: Zahlen, Datums-, Zeit- und Preisangaben, Satzzeichen und Emoticons (`:-)`). Auch HTML-Tags, die nicht zur Strukturierung dienen, sondern Dinge wie Bilder oder Formularfelder ausgeben, gehören dazu.

Manchmal reicht es nicht aus, nach Wortvorkommen an beliebigen Textstellen zu suchen, besonders wenn es sich um häufig verwendete Wörter handelt. Hier helfen *Strukturinformationen* weiter: Steht das Wort am Satz- oder Textanfang? Wird es durch HTML-Formatierungen wie z.B. Überschriften hervorgehoben? Ist es Teil eines Links?

In einige Bereichen ist die Verwendung von bestimmten Begriffen sogar vorgeschrieben. Für das Verfassen von Drehbüchern gelten beispielsweise Konventionen, welche Wörter in Regieanweisungen verwendet werden sollen. So steht »CONT.« für die Fortsetzung eines Gesprächs nach einem Einschub, der die Szenerie beschreibt oder »EXT.« bzw. »INT.« für die Festlegung, ob Schauplätze in oder außerhalb von Gebäuden liegen. Da der verwendete Korpus aus dem Internet stammt, kann man sich jedoch nicht darauf verlassen, dass jeder Hobby-Drehbuchautor sich daran hält – ein allgemeines Problem, auf das in Abschnitt 4.2.3 noch kurz eingegangen wird.

Abgeleitete Features

Bei einigen Genres darf die Anzahl von Types einer bestimmten Wortmenge nicht zu groß sein. In Interviews oder Drehbüchern gibt es nur eine begrenzte Zahl von Akteuren, weshalb auch nur wenige unterschiedliche Namen vorkommen, die dann allerdings sehr oft auftauchen. Auch bei Biografien und Porträts ist die Zahl der Namen beschränkt, da sie eine einzige Person beschreiben. Da diese Person entweder männlich oder weiblich ist, kann man zusätzlich beobachten, dass die Personalpronomen eines Geschlechts überwiegen (im Trainkorpus bei 95% um mindestens Faktor 4). Weitere Kombinationen von Features können der nachstehenden Liste entnommen werden.

4.2.2 Entscheidende Features je Genre

In folgender Aufzählung bedeutet *rel.* relativ zur Textlänge, *abs.* absolut und *neg.*, dass das Feature ein Ausschlussmerkmal für diese Klasse ist oder zumindest nicht zu häufig vorhanden sein darf. Da bei fast allen Genres die Textlänge berücksichtigt wird, wird sie hier nicht gesondert aufgeführt. Die Werte, die diese Features annehmen müssen, und ihre logischen Verknüpfungen können den einzelnen Programmen entnommen werden. Wenn die Gründe für die Auswahl der Merkmale nicht trivial sind, werden diese hier mit ▸ gekennzeichnet angegeben. Für alle Genres gilt: Verben und

Konjunktionen dienen zur Abgrenzung von Fließtext von Listen oder Ähnlichem, Adjektive (besonders positive und negative) sind ein Zeichen für den Grad der Sachlichkeit eines Textes. Je mehr von ihnen vorkommen, desto eher handelt es sich um einen erzählenden oder unterhaltsamen Text. »→Wortliste« bedeutet, dass die vollständige Auflistung der Wörter auf der beigefügten CD zu finden ist.

A.1 Kommentar

POS

- Adjektive *rel.*
- positive Adjektive (→Wortliste) *rel.*
- negative Adjektive (→Wortliste) *rel.*
- Verben *rel.*
- Verben in Past Tense relativ zu Verbanzahl ▸ in Past Tense verfasst
- Verben im Präsens relativ zu Verbanzahl
- Pronomen 1. Person *rel.* ▸ in Pamphleten viele (da persönlich)

Vokabular und Patterns

- Datum *rel., neg.*
- Kontraktionen (»n't«, »'ll«, »'d«, »'re«, »'ve«) *rel.* ▸ lockere Sprache
- Keywords für Rezensionen (Book, Film, CD, Disc) *rel., neg.*
- Kausalwörter (because, hence etc.→Wortliste) *rel.* ▸ Kommentare enthalten Begründungen

Kombinationen: Sprachstil

- Bösartig: Schimpfwörter + negative Adjektive - positive Adjektive + verallgemeinernde Wörter
- Argumentierend: 2 · Anzahl der Fragen + Kausal + Konditional + argumentierende Wörter + verallgemeinernde Wörter + Negationen

▸ Kommentare sind argumentierend oder Pamphlete

- lockere Sprache: Kontraktionen + verallgemeinernd + vage Wörter + informelle Wörter

wobei:

- verallgemeinernd (they, their, always, never, every, worst, worse, good, best, better, bad)
- vage (something, somehow, however, mysterious, perhaps)
- Schimpfwörter (drunk, wasteful, irresponsible, fool, stupid, crazy, laughable, faulty, bloated, coward, liar, blame, sick)
- argumentierend (But, finally, neveretheless, admit, actually, yes, despite, indeed, unless, either, question, aside, way, theory, furthermore)
- Konditional (would, should, could, will)
- informell (And, well, really, folks, great, only, this, dreamed, fear, feared, enough, guess und einzelne Wörter in Anführungszeichen)

A.2 Rezension

POS

- Adjektive *rel.*
- positive + negative Adjektive *rel.*
- Verben *rel.*
- Verben im Präsens rel. zur Anzahl der Verben
- Verben im Simple Past rel. zur Anzahl der Verben
- Verben im Präsens 3. Person, rel. zur Gesamtzahl der Verben und zu Verben im Präsens ▸ beschreibt Eigenschaften von Dingen/Ereignissen

- Pronomen 1. Person singular *rel.*
- Pronomen 3. Person singular, neut. *rel.*
- definite Artikel *rel.*
- indefinite Artikel *rel.*
- Verhältnis definiter zu indefiniten Artikeln ▸ mehr definite, da über bestimmte Dinge geredet wird.

Vokabular und Patterns

- Fragezeichen *rel., neg.*
- Schimpfwörter (→ A.1) *rel., neg.*
- vage (→ A.1) *rel., neg.* ▸ festes Urteil über das Thema, keine vagen Ausdrücke
- Deiktische Zeitangaben (yesterday, today, tomorrow, month, week) *rel., neg.*
- Keywords (book, film, movie, CD, album, anthology, designer, band, author, musician, restaurant, food, wine) *rel.* und *abs.*
- Kontraktionen *rel.* ▸ nur einige, keine allzu legere Sprache

Kombinationen

- Argumentierend (→ A.1)
- lockere Sprache (→ A.1)

A.3 Porträt, Nachruf, Würdigung

POS

- Personalpronomen 2. Person *rel., neg.* ▸ keine direkte Ansprache
- positive + negative Adjektive *rel.*

Vokabular und Patterns

- Maximum von »he« und »she« *rel.* und *abs.* ▸ berichtet über Person

Kombinationen

- Personalpronomen 1. Person relativ zur Anzahl der Anführungszeichen (direkte Rede) ▸ 1. Person nur in Zitaten, unpersönlich
- Verhältnis he/she ▸ eines überwiegt (ist ein Mann oder eine Frau das Thema?)
- Vorkommen des häufigsten Vornamens + 2, falls dieser in einer Überschrift am Textanfang vorkommt + Anzahl »he« bzw. »she« *rel.* ▸ in Bericht über eine Person sollte deren Namen in der Überschrift auftauchen
- Vorkommen des häufigsten Nachnamens + 1, falls dieser in einer Überschrift am Textanfang vorkommt + häufigster Vorname + Anzahl »he« bzw. »she« *rel.*
- Verhältnis von positiven zu negativen Adjektiven ▸ positiv überwiegt, oft freundliche Berichte

A.4 Glosse

POS

- Adjektive *rel.*
- positive + negative Adjektive *rel.*
- Verben *rel.*
- Pronomen 1. Person singular *rel.*
- Pronomen 2. Person *rel.*
- Pronomen 3. Person *rel.*

Vokabular und Patterns

- Keywords (column, columnist, irony, sarcasm, sarcastic, ironic) *rel.*
- Keywords für Rezensionen (→ A.1) *rel.*und *abs., neg.*

- Namen *rel.* ▸ entweder werden Namen genannt, oder der Leser wird mit »you« angesprochen
- Konditional (→ A.1) *rel.*
- vage (→ A.1) *rel.*
- Verben der Wahrnehmung (saw, recognize, wonder, notice, hear, see, note, spot, realize, perceive, glimpse, discover, overhear, listen, detect, observe) *rel.* ▸ Autor berichtet von seinen Erlebnissen/Gedanken
- Negationen *rel.* ▸ Stilmittel
- Kontraktionen *rel.* ▸ lockere Sprache
- Datum *rel., neg.*
- Ordinalzahlen (als Text, Zahl und römische Ziffern) *rel., neg.*
- Fragezeichen *rel.* ▸ Stilmittel (rhetorische Fragen)

Patterns in Struktur

- Begrüßung am Texanfang (→ F.1) *abs., neg.*
- Verabschiedung am Textende (→ F.1) *abs., neg.*

Kombinationen: Sprachstil

- lockere Sprache (→ A.1)
- Böse Sprache (→ A.1)
- argumentierend (→ A.1)

A.5 Interview

Vokabular und Patterns

- Keywords (interview, conversation, explained, told, discussed)
- FAQ-Keywords (FAQ, Q&A) *neg.*
- Wörter zur Zustimmung/Ablehnung (exactly, yes, indeed, no, yeah, not really, why not, OK, okay, don't think so, right) ▸ häufige Antworten des Interviewten

Patterns in Formatierung und Struktur

- Zeilen mit »?« oder »...« am Ende (Fragen) relativ zur Anzahl der Zeilen, *abs.*
- Regieanweisungen und Zeitangaben in Klammern bzw. Großbuchstaben (→ B.3) *neg.*

Kombinationen

- Zeilen mit Pronomen 1. und 2. Person (i, me, my, we, our, you, your) in Sätzen nach »Wort:« relativ zur Anzahl der Zeilen mit »Wort:« am Anfang ▸ oft »you« in Fragen, »I« in Antworten
- Anzahl der Interviewpartner (Wörter mit maximal 30 Zeichen am Zeilenanfang vor Doppelpunkt) ▸ nicht zu viele, oft nur zwei
- Anzahl der Interviewpartner mit mehr als einem Gesprächsbeitrag ▸ mindestens zwei

A.6 Nachricht

Struktur und HTML

- Anzahl der Sätze *abs.*
- Listenelemente (`<li>`-Tags) *abs., neg.*
- Überschriften *abs., neg.* ▸ berichtet von einem einzigen Ereignis, ist ungegliedert

POS

- Verben *rel.*
- Verben im Präsens relativ zur Anzahl der Verben
- Verben im Simple Past relativ zur Anzahl der Verben
- Verben in Verlaufsform relativ zur Anzahl der Verben
- Adjektive *rel.*

- positive + negative Adjektiv relativ zur Anzahl der Adjektive *neg.* ▸ neutral und sachlich
- Pronomen 1. Person *abs., neg.*
- Pronomen 2. Person *rel.,neg.*
- Konjunktionen *rel.*

Vokabular und Patterns

- Kausalwörter (→ A.1) *rel.,neg.* ▸ keine Gründe, sondern Fakten
- Namen *abs.* ▸ einige erlaubt, aber nicht zu viele
- Kontraktionen *rel., neg.*
- Ordinalzahlen *rel.* und *abs., neg.*
- Datum + Zeitwörter (months, years, weeks, now, yesterday, after, before, past, future, present) *rel.* ▸ Zeitpunkt des Geschehens muss genannt werden
- Fragezeichen *rel., neg.*
- Text-Formularelemente *abs., neg.*

Kombinationen

- Verhältnis Anführungszeichen zu Kontraktionen ▸ in direkter Rede sind sie erlaubt
- Verhältnis Anführungszeichen zu Pronomen 2. Person ▸ ebenso
- Differenz aus Pronomen 1. Person und Anführungszeichen ▸ ebenso
- Verhältnis Präsens zu Simple Past
- Verhältnis Präsens zu Verlaufsform

A.7 Feature

POS

- Verben *rel.*
- Verben im Präsens relativ zur Anzahl der Verben ▸ kurze Texte eher im Präsens
- Verben im Simple Past relativ zur Anzahl der Verben ▸ andere in der Vergangenheit
- Adjektive *rel.*
- positive + negative Adjektive *rel.* ▸ einige, da nicht ganz sachlich, aber nicht zu viele
- Pronomen 1. Person singular *rel.* ▸ alle Pronomen eher selten
- Pronomen 1. Person plural *rel.*
- Pronomen 2. Person *rel., neg.*
- Pronomen 3. Person *rel.*
- Konjunktionen *rel.*

Vokabular und Patterns

- Keywords für Porträts (born, died, saddened, Biography, Obituary, dies) *abs., neg.*
- typische Bigramme für wissenschaftliche Texte *abs., neg.*
- Maßeinheiten (→ C.3) *abs., neg.*
- argumentierende Wörter (→ A.1) *rel., neg.*
- verallgemeinernde Wörter (→ A.1) *rel., neg.*
- informelle Wörter (→ A.1) *rel., neg.*
- Namen *rel.* ▸ nicht zu viele
- Vergangenheits-Keywords (»Century«, »Decade« und Jahreszahlen vor 1980) *rel., neg.*
- Kontraktionen *rel., neg.*
- Datum *rel.*
- Fragezeichen *rel., neg.*
- Text-Formularelemente *abs., neg.*

Patterns in Struktur

- Keywords für andere Genres am Textanfang (Interview, Review, Annual, Monthly) *abs.*, *neg.*

Kombinationen

- Namen + Pronomen 3. Person als Kennzeichen für Personen-Nennung ▸ nicht zu viele
- Verhältnis Präsens zu Simple Past

A.8 Reportage

POS

- Verben *rel.*
- Verben im Präsens relativ zu Anzahl der Verben
- Verben im Simple Past relativ zu Anzahl der Verben
- Adjektive *rel.* ▸ bildhafte Sprache, verwendet viele Adjektive
- positive + negative Adjektive *rel.*
- Pronomen 1. Person *rel.* ▸ oft Bericht über selbst erlebtes…
- Pronomen 3. Person *rel.* ▸ … und über andere Personen
- Konjunktionen *rel.*

Vokabular und Patterns

- Keywords für Porträts (→A.7) *abs.*, *neg.*
- typische Bigramme für wissenschaftliche Texte (→Wortliste) *abs.*, *neg.*
- argumentierende Wörter (→A.1) *rel.*, *neg.*
- verallgemeinernde Wörter (→A.1) *rel.*, *neg.*
- informelle Wörter (→A.1) *rel.*, *neg.*
- Namen *rel.* ▸ Bericht über Personen und Lebewesen oder Länder und Städte
- Länder und Städte *rel.*
- »the« gefolgt von Lebewesen (Berufe, Verwandschaftsbezeichnungen, Tiere →Wortliste) *rel.*
- Vergangenheits-Keywords (→A.7) *rel.*, *neg.*
- Kontraktionen *rel.*, *neg.*
- Datum *abs.* und *rel.*
- Fragezeichen *rel.*, *neg.*
- Text-Formularelemente *abs.*, *neg.*

Kombinationen

- Namen + Pronomen 3. Person
- Verhältnis Präsens zu Simple Past

B.1 Gedicht

Struktur

- durchschnittliche Zeilenlänge des längsten Gedicht-Blocks
- Länge diese Blocks in Zeilen relativ zur gesamten Zeilenzahl und abs.
- Länge diese Blocks in Zeichen *rel.*

wobei ein Gedicht-Block definiert wird als:

- aufeinanderfolgender Zeilen, die 5 bis 70 Zeichen lang sind und nicht zu viele Tags und Sonderzeichen enthalten
- enthält wenige Namen, Städte, Länder, Titel (Mr, Mrs etc.) und Firmennamen (Company, Associates etc.→Wortliste)
- enthält wenige Anführungszeichen
- enthält wenige Zahlen

- enthält wenige Doppelpunkte am Zeilenende
- wenn der Block in <pre>-Tags steht, so muss er den kompletten Bereich ausfüllen

Vokabular und Patterns

- Kapitelüberschriften (»Chapter«) *abs.*, *neg.*

Patterns in Block

- Wörter in Großbuchstaben rel. zur Textlänge des Blocks, *neg.*
- Sonderzeichen rel. zur Textlänge des Blocks, *neg.*
- Zahlen rel. zur Textlänge des Blocks, *neg.*
- Doppelpunkte am Zeilenende rel. zur Textlänge des Blocks, *neg.*
- typische in Sätzen vorkommende Wörter (Personal- und Fragepronomen, Konjunktionen, Hilfsverben, Konditionalwörter) *abs.* ▸ um sicherzustellen, dass es ein Text ist

B.2 Roman

Struktur

- Satzanzahl
- durchschnittliche Satzlänge ▸ eher lange und komplexe Sätze

POS

- Adjektive *rel.*
- Verben *rel.*
- Verben im Präsens *rel.*
- Verben im Simple Past *rel.*
- Personalpronomen *rel.*
- Konjunktionen *rel.*

Vokabular und Patterns

- Kommas *rel.* ▸ komplexe Sätze
- Zahlen *rel., neg.*
- Anführungszeichen *rel.* ▸ verwenden relativ viel direkte Rede
- Synonyme für Sprechen (→ Wortliste) ▸ als Einleitung der direkten/indirekten Rede
- altertümliche Ausdrücke *rel.*

Patterns in Formatierung und Struktur

- Doppelpunkt nach ersten 30 Zeichen einer Zeile *rel.*

Kombinationen

- Konjunktionen + Kommas ▸ als Maß für Satzkomplexität
- Verhältnis Pronomen 1. Person zu allen Personalpronomen ▸ viele in Dialogen
- Verhältnis positiver zu negativen Adjektiven ▸ relativ ausgewogen
- Anzahl Doppelpunkte abhängig von Anzahl der Synonyme für Sprechen ▸ wenn viele Synonyme vorkommen, sind mehr Doppelpunkte erlaubt
- Namen + »the« gefolgt von Lebewesen (→ A.8) *rel.* ▸ Handelnde Personen
- Pronomen + Namen + »the« gefolgt von Lebewesen + Anführungszeichen *rel.* ▸ Zeichen für Dialoge und Handelnde

B.3 Drama

Struktur und HTML

- <pre>-Tags

Vokabular und Patterns

- Zahlen *rel., neg.*
- Namen *rel.*
- auf »-ing« oder »-ly« endende Wörter (z.B. »softly«, »whispering«) oder Regieanweisungen (fade, dissolve, pause, blackout, black screen) in Klammern
- mit »Wort:« beginnende Zeilen relativ zur Zeilenzahl

Patterns in Formatierung und Struktur

- Anzahl typischer Regieanweisungen zu Sprechertexten (CONT., CONT'D, VO, V.O., CONTINUED) in Großbuchstaben
- Anzahl typischer Zeit- und Ortsangaben in Regieanweisungen (AFTERNOON, DUSK, DAWN, SUNNY, RAINY, SNOW, LATER, SIMULTANEOUSLY, EXT., INT.) in Großbuchstaben
- Anzahl Keywords (scene, act, cast, roles, characters, script, drama, theater) in den ersten Tausend Zeichen

C.1 Wissenschaftlicher Bericht

Vokabular

- Anzahl typischer Bigramme mit »we« oder »our« (→ Wortliste)

Patterns in Formatierung und Struktur

- Anzahl der sehr häufig und fast ausschließlich in wissenschaftlichen Texten verwendeten Überschriften (Abstract, Synopsis, Acknowledgments, References, Introduction, Discussion oder Conclusion in `<h>`, `<b>`, `<font>` oder `<center>`)
- Anzahl der häufig in wissenschftlichen Texten verwendeten Überschriften (Results, Appendix, Summary, Evaluating, Evaluation, Bibliography, Biography, Footnotes, Synthesis, Description, Suggested Reading, Ressources, Previous Studies)

C.2 Erklärung

POS

- Verben *rel.*
- Verben im Präsens relativ zur Anzahl der Verben
- Verben im Simple Past relativ zur Anzahl der Verben
- Adjektive *rel.*
- positive + negative Adjektive *rel.* ▸ eher wenige, da sachlicher Text
- Pronomen 1. Person plural *rel., neg.*
- Pronomen 3. Person singular neut. *rel.*
- Konjunktionen *rel.*
- definite Artikel *rel.*

Vokabular und Patterns

- vage (→ A.1) *rel., neg.*
- typische Bigramme für wissenschaftliche Texte (→ Wortliste) *abs., neg.*
- Namen *rel.*
- Maßeinheiten (→ C.3) *rel., neg.*
- Vergangenheits-Keywords (→ A.7) *rel.* ▸ geschichtliche Erklärungen
- Deiktische Zeitangaben (→ A.2) *rel., neg.* ▸ unpersönlich, wenig Involviertheit
- Kontraktionen *rel., neg.*
- Kontraktionen ohne »n't« *rel., neg.*
- Zahlen *rel.* ▸ einige (Größenangaben, Jahreszahlen etc.), aber nicht zu viele

- Ordinalzahlen *rel., neg.*
- Datum *rel., neg.*
- Fragezeichen *rel., neg.*

Patterns in Formatierung und Struktur

- Überschriften für wissenschafltiche Texte (→C1) *abs., neg.*
- Datum, dass nicht in Links steht *rel.*und *abs., neg.*

Kombinationen

- Namen + Pronomen 3. Person
- Verhältnis definiter zu indefinten Artikeln ▸ mehr bestimmte (sachlicher Text zu einem Thema)
- Namen abhängig von Vergangenheits-Bezeichnern ▸ historische Schilderungen, mehr erlaubt

C.3 Anleitung, Rezept

POS

- Pronomen 1. Person singular *rel.*
- Pronomen 1. Person plural *rel.*
- Pronomen 2. Person singular *rel.*
- Pronomen 3. Person singular *rel., neg.*
- Konjunktionen *rel.*
- Sätze, die mit Verb beginnen *rel.* ▸ Befehlssätze, Anweisungen
- Wörter, die auf »-ing« enden *rel.*
- Verben in Past Tense *rel.* und *abs.*

Vokabular und Patterns

- Maßeinheiten (Tablespoon, Cup, Teaspoon, oz, tbsp, tsp)
- Ordinalzahlen (als Text und Zahl, ohne römische Ziffern), inklusive <o1>-Einträge *rel.*
- Namen ▸ nicht zu viele, handelt nicht von Personen
- Text-Formularfelder, *neg.*

Patterns in Struktur

- Überschrift-Keywords (How To, Tutorial, Guidelines, Using, Recipes)

Kombinationen

- Sätze, die mit Verb beginnen + Pronomen 1. Person Plural und 2. Person ▸ Möglichkeiten, wie Anweisungen formuliert werden können
- Pronomen 1. Person, 2. Person oder Maßeinheiten ▸ entweder »dann habe ich«, »du musst« oder Rezepte ohne Subjekt
- »ing«-Form, Maßeinheiten oder Ordinalzahlen ▸ mögliche Formulierungsarten

C.4 FAQ

Vokabular und Patterns

- Keywords (FAQ, Q&A, Frequently Asked Questions) *rel.*
- Text-Formularfelder *abs., neg.*

Patterns in Struktur

- Keywords im Text ohne Links *rel.*
- Anzahl der Fragen (Sätze am Zeilenanfang, die kürzer als 200 Zeichen sind und mit Fragezeichen enden) relativ zur Zeilenanzahl

Kombinationen

- Anzahl der Fragen mit typischen Fragewörtern (who, when, where, what, how, why, which,

whom, can I, do I, may I, could I) relativ zur Anzahl der Fragen
- Anzahl der Gesprächsteilnehmer (Wörter vor Doppelpunkt, die mindestens zwei mal vorkommen und keines der Ausschluss-Keywords (Q, A, Question, Answer, Subject, http, From, Date) sind *abs.*, *neg.*

C.5 Glossar

Struktur

- Länge einer Liste mit alphabetisch sortierten Wörtern
- Fehler in Sortierung relativ zur Listenlänge

Vokabular und Patterns

- Keywords (Glossary, Definition, Lexicon, Reference, Dictionary, Terms; aber nicht »usage terms« oder »terms of usage«)
- Alphabet (wobei Buchstaben-Bereiche wie A-E auch erkannt werden) ▸ typische Navigation mit Links für jeden Buchstaben

C.6 Zweisprachiges Wörterbuch

Vokabular

- Verhältnis Namen (aufeinanderfolgende Großgeschriebene Wörter) zu Wörtern im Gerneral English (→ Wortliste)
- Verhältnis der Wörter nicht in General English + Ländernamen + Codewörtern + Namen zu Wörtern im Gerneral English

C.7 Präsentation

POS

- Verben *rel.*
- Verben im Präsens relativ zur Anzahl der Verben
- Verben im Präsens 3. Person relativ zur Anzahl der Verben
- »are« gefolgt von Past Participle *rel.*
- »has been« *rel.*
- Pronomen 1. Person singular *rel., neg* ▸ unpersönlich
- Pronomen 1. Person plural *rel.*
- Pronomen 3. Person singular *rel., neg.* ▸ kein Bericht über Dritte
- Pronomen 3. Person plural *rel.,neg.* ▸ keine Verallgemeinerungen
- Konjunktionen *rel.*

Vokabular und Patterns

- vage (→ A.1) *rel., neg.* ▸ stellt ganz bestimmt Fakten und Werbung dar
- typische Bigramme für wissenschaftliche Texte *abs.*, *neg.*
- Maßeinheiten (→ C.3) *rel.*, *neg.*
- Ordinalzahlen *rel., neg.*
- Datum *rel.,neg.*
- Fragezeichen *rel.,neg.*

Patterns in Formatierung und Struktur

- Keywords (welcome, new, philosophy, information, mission, founded, goal, innovative, innovation, global) am Textanfang
- Keywords als Bildname ▸ kommt häufig vor, da solche Seiten oft sehr grafisch sind

Kombinationen

- Verhältnis definiter zu indefinten Artikeln ▸ spricht über bestimmte Dinge
- 1. Person plural, »has been«, »are» + Past Participle oder Verben in 3. Person singular ▸ verschiedene Stile: »we are the most fabulous«, »this company has been«, »are advised« oder »this product is great«

C.8 Statistik

Struktur und HTML

- erstellt mit Excel (ein Tag enthält `class=xl...`)

Vokabular und Patterns

- Zahlen *abs.* und *rel.*
- Prozentzahlen *abs.* und *rel.*

Patterns in Formatierung und Struktur

- nach amerikanischem Standard formatierte Zahlen in Tabellenzellen *abs.*
- aufeinanderfolgende solche Patterns *abs.*, *neg.*

C.9 Code

Struktur und HTML

- Codetags: `<pre>`, `<xmp>`, `<code>`, `<samp>`; `<font>`-Tags mit Angabe der Monospace-Schriftart »Courier« oder CSS-Klasse mit Namen »preformat«, »pre« oder »code« *abs.*
- aufeinanderfolgende Zeilen, die auf eines der Zeichen `; > { }` enden, wobei Zeilen mit Kommentaren (`/*comment */`, `//comment` oder `#comment`) nicht betrachtet werden *rel.*

Vokabular und Patterns

- Keywords in Porgrammiersprachen *rel.* (→ Wortliste)
- typische Variablennamen wie bla1, blaBla, $bla, bla_bla *rel.*

D.1 Gesetz

POS

- Pronomen 2. Person *rel., neg.* ▸ keine Ansprache des Lesers
- Adjektive *rel., neg* ▸ nüchtern und sachlich

Vokabular und Patterns

- Keywords (Amendment, Constitution, Act, Proclamation, Statutes, Code, Contract, Bill, Rules) *rel.*
- Namen *rel., neg.*
- Zahlen *rel.* ▸ Nummerierung der Gesetze, Abschnitte...

Patterns in Formatierung und Struktur

- Ordinalzahlen am Zeilenanfang, inklusive `<ol>`-Einträge + Article/Section gefolgt von Nummerierung (Kleinbuchstaben, Ordinalzahlen) am Zeilenanfang ▸ ist streng gegliedert

D.2 Offizieller Bericht

POS

- Verben *rel.*
- Verben im Präsens relativ zur Anzahl der Verben ▸ eher wenige, da...
- Verben im Simple Past relativ zur Anzahl der Verben ▸ ... Bericht über Vergangenes
- Pronomen 1. Person singular *rel., neg.*
- Pronomen 2. Person *rel., neg.*

- Pronomen 3. Person singular *rel., neg.*

Vokabular und Patterns

- Schimpfwörter (→ A.1) *rel., neg.*
- vage (→ A.1) *rel., neg.*
- Konditional (→ A.1) *rel., neg.* ▸ Faktendarstellung, keine Vermutungen
- typische Zeitangaben bei Veröffentlichung von Berichten (Year, Annual, Month, Monthly) *rel.*
- Kontraktionen *rel.*
- Preisangaben (→ E.2) *rel.* ▸ Zahlen zum Geschäftserfolg
- Ordinalzahlen *rel.* und *abs.* ▸ gegliedert
- Datum *rel.* ▸ enthält detaillierte Zeitangaben
- Fragezeichen *rel., neg.*

D.3 Meeting Minutes

POS

- Verben *rel.*
- Verben im Präsens relativ zur Anzahl der Verben
- Verben im Simple Past relativ zur Anzahl der Verben ▸ Bericht über Vergangenes
- Pronomen 1. Person singular *rel., neg.*
- Pronomen 2. Person *rel., neg.*
- Pronomen 3. Person singular *rel., neg.*

Vokabular und Patterns

- Schimpfwörter (→ A.1) *rel., neg.*
- vage (→ A.1) *rel., neg.*
- informelle Sprache (→ A.1) *rel., neg.*
- Konditional (→ A.1) *rel., neg.*
- Kontraktionen *rel., neg.*
- Ordinalzahlen *abs.* und *rel.* ▸ gegliedert
- Datum *rel.*
- Fragezeichen *rel., neg.*
- Text-Formularfelder *abs., neg.*

Kombinationen

- lockere Sprache (→ A.1), *neg.*
- Keywords (Minutes, Meeting) + 1 falls eines der Wörter members, present, absent, location, met, attendance, approval, called oder order vorkommt, -1 falls »Report« vorkommt ▸ diese Wörter kommen auch in anderen Texten vor, deswegen geringe Gewichtung

E.1 Personen-Verzeichnis

sucht listenähnliche Strukturen mit Namen, Firmen, Städten oder Ländern und bestimmt:

- Anzahl der Zeilen *abs.* und *rel.*
- Anzahl der sortierten Zeilen *abs.* und *rel.*
- Sortierfehler relativ zur Zeilenzahl, neg.

E.2 Katalog

- Preisangaben, bestehend aus Zahl und Währung (Dollar, Euro und Pfund) *rel.* und *abs.*
- Text-Formularelemente *abs.* ▸ Such- oder Bestellfelder

E.3 Ressourcen

a) Literatur

- Gesamtlänge aller Reference-Patterns (enthält Datum oder ISBN-Nummer und Text) *rel.*
- darin enthaltene Nomen (relativ zu Nomen+Verben+Adjektive und *abs.*) ▸ Buchtitel, Namen
- darin enthaltene Verben (relativ zu Nomen+Verben+Adjektive) ▸ eher wenige
- darin enthaltene Uhrzeitabgaben *abs.*, *neg.*
- darin enthaltene Wochentagen *abs.*, *neg.*

b) Links

- Anzahl der Zeilen mit ähnlicher Struktur (z.B. immer »Link - Datum - Text«) in Tabellen, Listen u.ä., die mit Links anfangen oder enden; *rel.*, *abs.* und relativ zum längsten Textblock ohne Links
- Länge des längsten zusammenhängenden Blocks ohne Links *abs., neg.* ▸ um Seiten mit langen Navigationslisten asuzuschließen
- Text-Formularfelder, *neg.*

E.4 Timelines

sucht listenähnliche Strukturen die mit Zeitangabe in einem bestimmten Format beginnen (Jahreszahl, Jahr-Monat, Jahr-Monat-Tag oder Uhrzeit) und bestimmt:

- Länge der Liste
- Anzahl der Fehler *abs.* und relativ zu Zeilenzahl der Liste, *neg.*
- Anzahl der Änderungen in Sortierung: Zurücksetzen auf frühere Zeitangabe oder Ändern der Sortier-Reihenfolge (auf- oder absteigend?), *neg.*
- $(1 - \text{Fehler}^2/\text{Listenlänge}^2) \cdot \text{Listenlänge/Textlänge}$ ▸ bei absolut und relativ langen Listen sind mehr Fehler erlaubt

F.1 Brief, Mail, Rede

POS

- Verben *rel.*
- Verben im Präsens 3. Person relativ zur Anzahl der Verben
- »are« gefolgt von Past Participle *rel.*
- »has been« *rel.*
- Pronomen 1. Person singular *rel.* ▸ persönlich
- Pronomen 2. Person *rel.* ▸ persönlich
- Pronomen 3. Person plural *rel.*
- Pronomen 3. Person gesamt *rel.*

Vokabular und Patterns

- Keywords (writing, lines, readers, written, editor, response, received, saying) *abs.*
- deiktische Wörter (here, now, today) *rel.* ▸ persönliche Kommunikation, hohe Involviertheit
- Maßeinheiten (vgl. C.3) *rel.*, *neg.*
- Zahlen *rel., neg.*
- Fragezeichen *rel.*

Patterns in Struktur

- Begrüßung am Texanfang (Dear, Hi, Hello, Madame, Mister, Good morning/afternoon/evening) *abs.*
- offizielle Begrüßung am Texanfang (To the Editor, To whom it may concern, Open Letter, Address on, Madame, Mister, President) *abs.*

- Verabschiedung am Textende (Thank you, Sincerely, Regards, Best Regards, Yours Sincerely, Yours faithfully, Best Wishes, Best to all, Signed) *abs.*
- Keywords für Anleitungen und FAQs am Textanfang (how to, tutorials, guidelines, using, recipes, FAQ) *abs.*, *neg.*

Kombinationen

- Ausrufezeichen + Fragezeichen *rel.* ▸ lebendige Sprache, Stilmittel

F.2 Forum

Struktur und HTML

- Verhältnis dargestellter Text zu HTML-Code (Content-to-Code-Ratio, CCR) ▸ gering durch Verwendung von Foren-Software
- Anzahl Überschriften ▸ nur wenige, ist ungegliedert

Vokabular und Patterns

- Emoticon-Bilder: `src` im `<img>`-Tag enthält »smilies« oder »emoticons«; Bild ist im GIF-Format
- Emoticons (→ Wortliste) *rel.*
- Acronyme (→ Wortliste) *rel.*
- Keywords (Re:, Posts:) *rel.*
- Blog-Keywords in Links (→ F.3) *abs.*, *neg.*
- Zeichen-Wiederholungen (z.B. Haaaaallo!!!!)
- Datum *rel.* ▸ jeder Beitrag ist mit Datum gekennzeichnet
- File-Links (FTP-Protokoll oder Dateiendung pdf, zip, ps, ppt, gz, doc), *neg.*

Kombination

- gewichtete Summe aus Emoticons, Acronymen, Emoticon-Bildern, Zeichen-Wiederholungen und Keywords

F.3 Blog

Struktur und HTML

- `<link>`-Tag für RSS *abs.*

Patterns in Struktur

- Blog-Anbieter-Keywords in Links (Movable Type, Wordpress, HaloScan)
- Keywords in Links (Comment, Trackback, Permalink, RSS)
- »posted:« nach schließenden Tags

F.4 Formulare

- Anzahl der Text-Input-Felder im längsten Formular des Dokuments
- analog dazu Anzahl der Select-Felder
- Gesamtlänge aller Formulare (HTML) *rel.*

G Nichts

- Fehlermeldungen etc. (error, file not found, index of , page not found, could not be found, 404 etc.) *abs.*
- Fehlermeldungen von Skriptsprachen oder Webhostern (»Stack Trace«, »Seeing this instead of the website you expected«, »web server [...] problem«) *abs.*
- Zahlen *abs., neg.*
- Währungen *abs., neg.*

4.2.3 Probleme

Texte, die keine besondere Struktur und auch kein spezifisches Vokabular aufweisen, lassen sich oft nur mit Hilfe von POS-Merkmalen oder durch die Art der verwendeten Sprache unterscheiden. Allerdings stößt man selbst dabei auf Grenzen, da die Unterschiede im Stil zweier Autoren größer sein können, als die zwischen Genres. Argamon et al. konnten zeigen, dass sich der Stil von Männern und Frauen stärker unterscheidet, als der von Nonfiction- und Fiction-Texten. Ein Beispiel ist die Verwendung von Proper Nouns, die allgemein seltener von Frauen als von Männern und häufiger in Non-Fiction als in erfundenen Texten auftreten. Die mittlere Anzahl dieser Wörter bei von Frauen verfassten Non-Fiction-Texten liegt mit 213/10 000 beinahe genau in der Mitte zwischen derjenigen in Fiction-Texten von Männern (226) und Frauen (198). Ähnliches gilt für Pronomen in der 1. Person plural. [ARG] Möglicherweise könnte die Bestimmung des Geschlechts des Autors auch die Genreklassifikation verbessern, da spezifische Schreibstile berücksichtigt werden könnten.

Außerdem gibt es bei einigen Textarten nur minimale Differenzen. Glossen und Kommentare geben beide auf eher lockere Art die Meinung des Verfassers wieder, der Unterschied liegt nur in der *Art* des Inhalts. Informationen (in Form von Themenlisten o.ä.) zum Grad dessen Aktualität und des öffentlichen Interesses wären hier notwendig. Auch weitergehende Analysen von häufig verwendeten Phrasen können hilfreich sein. Einfache Part-of-Speech- oder Sprachstil-Merkmale reichen jedenfalls nicht aus.

Listen

HTML bietet eigene Konstrukte an, um Listen zu erzeugen. Man hat die Wahl zwischen geordneten und ungeordneten sowie Definitionslisten. Im Prinzip dürfte die Erkennung von Listen also kein Problem darstellen – in der Realität leider schon. Die Autoren von Websites verwenden beliebige andere Strukturierungs- und Formatierungsmittel, um Aufzählungen darzustellen: Tabellen, Absätze, Layer (`<div>`), `<br>`-Zeilenumbrüche und Auszeichnungen (fett, Großbuchstaben), oder im schlimmsten Fall preformatierten Text, in den durch Leerzeichen und Umbrüche eine (optische) Struktur gebracht wurde.

Um Listen zu finden, muss daher das Dokument auf Regelmäßigkeiten und Muster analysiert werden. Hierzu werden die oben genannten Pseudo-Listenelemente der Reihe nach durchlaufen und darin das gewünschte Muster, also Namen, Datumsangaben etc., gesucht. Wenn in der ersten der untersuchten Strukturen (z.B. Listen) keine ausreichende Anzahl dieser Elemente gefunden wird, wird die gleiche Prozedur auf die nächste Möglichkeit (z.B. Tabellen) angewandt. Die Reihenfolge wird aus den Häufigkeiten der jeweilgen Strukturierungs-Methode im Trainingskorpus bestimmt. Bei Personenlisten sind die durchsuchten Elemente zuerst Definitionslisten, dann geordnete oder ungeordnete Listen, anschließend Tabellen gefolgt von Texten in fett oder Großbuchstaben und zu guter Letzt Wörter direkt am Zeilenanfang. Glossarbegriffe stehen in Definitionslisten, fett oder groß geschriebenen Begriffen gefolgt von einem Doppelpunkt oder Ähnlichem, oder in normalen Listen.

Bei Timelines gibt es die zusätzliche Schwierigkeit, dass das Datum in unterschiedlichen Formaten angegeben werden kann. Deswegen werden alle Daten zuerst in ein leicht sortierbares normiertes Format konvertiert (wie 2005-12-24, 12-24, 2005 oder 08:45). Anschließend werden Definitionslisten, ungeordnete Listen (geordnete würden hier keinen Sinn ergeben), Tabellen und Zeilenanfänge

duchsucht und für jede der vier Datumsarten die Vorkommen gezählt, bis deren Maximum für eine Struktur einen bestimmten Schwellenwert erreicht.

Die so gefundenen Listen werden anschließend daraufhin geprüft, ob sie sortiert sind. Glossare und Personenlisten beginnen mit dem alphabetisch kleinsten Element, wobei letztere nach Vor- oder Nachname sortiert werden können. Bei Timelines ist die Sortierung nicht vorgegeben sondern muss erst noch durch eine Analyse der ersten 20 Einträge bestimmt werden. Treten später mehr als drei Verletzungen dieser Regel auf, so wird diese Annahme revidiert und die erwartete Reihenfolge umgedreht. Bei allen drei Listen dürfen einige Fehler auftreten, da der Algorithmus möglicherweise falsche Elemente mit in die Liste aufgenommen hat (z.B. Zwischenüberschriften), dem Verfasser Fehler beim Ordnen unterlaufen sind oder mehrere sortierte Folgen hintereinander auftauchen können, beispielsweise wenn ein Veranstaltungsprogramm für mehrere Tage dargestellt wird.

Auch sonst muss man stets im Hinterkopf behalten, dass die Dokumente zum Teil von Laien erstellt oder von nicht darauf spezialisierten Programmen generiert worden sind. Man findet veraltete Tags, nicht geschlossene Formulare, mehrere `<head>`-Elemente, falsch beendete Kommentare, sinnlose Verschachtelungen etc. All diese Fehler müssen bei der automatischen Extraktion der Merkmale berücksichtigt werden. Wünschenswert wäre eine Art Reinigungsprogramm, welches mangelhaften HTML-Code repariert. Bei einer kurzen Recherche konnte ich jedoch keines finden, und auch den Versuch nebenbei ein solches Programm zu schreiben, habe ich (überwältigt durch die Unmenge an unterschiedlichen Fehlern) schnell wieder aufgegeben.

5. Klassifikation

Ausgehend von den ermittelten Features soll nun ein Verfahren entwicklelt werden, das unbekannte Texte automatisch einem Genre zuordnet. Dazu wurden die Ergebnisse der Einzel-Klassifikatoren auf verschiedene Arten kombiniert. In einer weiteren Versuchsreihe wurden unter Verwendung sämtlicher Features (vgl. 9.3) mehrere automatische Verfahren getestet.

5.1 Klassifizierungs-Algorithmen

Zur automatischen Klassifizierung von Datensätzen gibt es mehrere Ansätze, von denen einige in diesem Kapitel vorgestellt und ihre Eignung für die Verarbeitung von Texten bewertet werden.

Naive Bayes

Bayes-Klassifikatoren bestimmen die Wahrscheinlichkeit, mit der ein Objekt (bzw. sein Merkmalsvektor) zu einer Klasse gehört. Bei der Klassifikation wird nach dem *Maximum Likelihood-Prinzip* die mit dem höchsten Wert gewählt. Zur Ermittlung der Wahrscheinlichkeiten wird der *Satz von Bayes* verwendet, der aus der Wahrscheinlichkeit des Merkmals M (z.B. die Anzahl der Adjektive), derjenigen von M innerhalb einer Klasse (Adjektive in Romanen) und derjenigen der Klasse (Wahrscheinlichkeit, dass ein Text in Roman ist) berechnet, wie wahrscheinlich eine Klassenzugehörigkeit bei gegebenem Merkmal ist (Mit welcher Wahrscheinlichkeit ist der Text ein Roman, wenn x Adjektive vorkommen?). Gibt es mehrere Merkmale, so ist M ihre Kombination. Die benötigten Werte können aus den Trainingsdaten gewonnen werden.

$$P(C_1 | M) = P(M | C_1) \cdot P(C_1) / P(M)$$

Formel 5.1: Satz von Bayes. C ist die Klasse, M das Merkmal [KDD, S. 107]

Bei hochdimensionalen Merkmalsvektoren gibt es sehr viele Kombinationsmöglichkeiten der einzelnen Merkmale: n Merkmale mit jeweils x möglichen Werten führen zu x^n Kombinationen. Der Trainingskorpus müsste sehr groß sein, um eine ausreichende Anzahl von Objekten für jede Kombination zu enthalten. Der Naive Bayes-Klassifikator löst dieses Problem, indem er annimmt, dass die Merkmale nicht voneinander abhängen, also statistisch unabhängig sind. Die Wahrscheinlichkeit, dass beide Merkmale erfüllt sind ergibt sich dann aus dem Produkt der Einzelwahrscheinlichkeiten.

Falls die Unabhängigkeits-Bedingung nicht erfüllt ist, bedeutet dies allerdings noch nicht, dass das Verfahren komplett versagt [KDD, S. 113]. Zudem kann man das Problem dadurch verringen, dass man dem Klassifikator geeignete Kombinationen von Features zur Verfügung stellt, die die Abhängigkeiten berücksichtigen.

Entscheidungsbaum (C4.5)

Entscheidungsbäume sind der Klassifikation durch den Menschen am nähesten. Sie entsprechen einer Reihe von wenn-dann-Tests, die sich durch anschauliche Sätze ausdrücken lassen, zum Beispiel: »wenn der Text sehr viele Namen enthält, dann handelt es sich um eine Personenliste« oder »wenn der Text eine Listenstruktur aufweist (Bedingung 1) und die einzelnen Punkte alphabetisch geordnet sind (Bedingung 2), dann ist es ein Glossar« (Abb. 5.1). Jede Bedingung entspricht einem

Knoten in einer Baumstruktur, jede Klasse einem *Blatt*. Zur Klassifikation wird der Baum von der Wurzel an durchlaufen bis ein Blatt erreicht ist.

Es gibt mehrere Verfahren, um Entscheidungsbäume zu generieren. Verwendet wurde der von WEKA zur Verfügung gestellte C4.5 bzw. J48.

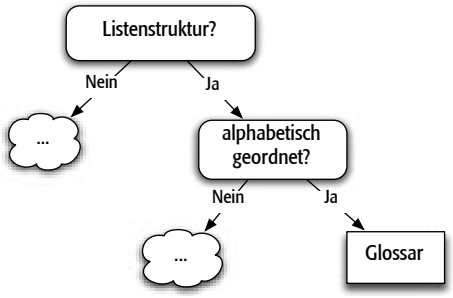

Abbildung 5.1: Entscheidungsbaum

Der Algorithmus teilt die Objektmenge so lange auf, bis sich in einem Knoten nur noch Objekte einer Klasse befinden – dieser wird dann zum Blatt (Top-Down Induction of Decision Trees). Das Merkmal nach dem jeweils gesplittet wird, wird so gewählt, dass der Informationsgewinn (Information Gain) möglichst groß ist. Dieser berechnet sich aus der Differenz zwischen Entropie des Ursprungsknoten und den Entropien der entstandenen Knoten. Die Entropie ist ein Wert für die Reinheit einer Menge. Sie erreicht ihr Minimum, wenn alle Objekte in einem Knoten aus der selben Klasse stammen.

Bei derartig generierten Entscheidungsbäumen kann allerdings *Overfitting* auftreten, falls zufällige Phänomene oder Fehler in der Trainingsmenge zur Auswahl der Bedingungen verwendet werden. Die Güte der Klassifikation nimmt auf Grund solcher falschen Entscheidungsregeln ab. Um Overfitting zu vermeiden, verwendet J48 *Reduced Error Pruning*. Dazu wird die Objektmenge in Trainings- und Testdaten unterteilt. Aus dem mit Hilfe der Trainingsobjekte erstellten Baum wird testweise jeder Knoten entfernt und geprüft, ob sich dadurch Verbesserungen für die Testmenge ergeben. Falls dies der Fall ist, wird der Knoten gelöscht. [GRI]

Nearest Neighbours

Beim K-Nearest-Neighbour-Verfahren wird ein Objekt derjenigen Klasse zugeordnet, der es im Merkmalsraum am nächsten ist. Der Abstand zwischen zwei Objekten bestimmt sich dabei aus den Differenzen der einzelnen Attribute. Im einfachsten Fall (K=1) wird nur der nächste Nachbarpunkt betrachtet, sonst die K nächsten. Dabei kann man die einzelnen Nachbarn nach ihrem Abstand gewichten; nähere haben einen größeren Einfluss. Der optimale Wert für K kann durch Kreuzvalidierung auf den Trainingsdaten automatisch bestimmt werden. Je größer der Wert ist, desto langsamer wird die Einordnung der neuen Objekte. Dafür sind die Ergebnisse oft sehr gut.

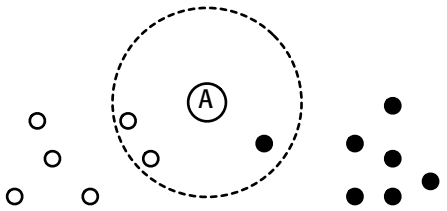

Abbildung 5.2: k-Nearest-Neighbour-Klassifikation im 2-dimensionalen Raum, k = 3. Das neue Objekt A wird der weißen Klasse zugeordnet, da 2 der 3 nähesten Punkte weiß sind.

Support Vector Machines

Support Vector Machines (SVM) teilen Daten in zwei Klassen, indem sie eine geeignete trennende Hyperebene im Feature-Raum berechnen. Diese hat die Eigenschaft, dass die Punkte auf beiden Seiten möglichst weit von ihr entfernt sind und dass sie beim Einfügen neuer Daten nur mir geringer Wahrscheinlichkeit geändert werden muss (*Maximum Margin Hyperplane*). Wenn man die Klasse eines Objekts bestimmen möchte, muss man lediglich bestimmen, auf welcher Seite der Ebene es liegt. Da Objekte nicht immer linear trennbar sind, kann man sogenannte *Kernel-Funktionen K* verwenden, welche die Daten in einen anderen Raum höherer Dimension transformieren. Wenn die Objekte (mit zwei Merkmalen) z.B. durch eine Parabel separierbar wären, wählt man als Kernel

$$K(x_A, x_B) = (x_A, x_B)^2$$

Formel 5.2: Quadratischer Kernel

Dies gilt für eine Transformation der Objekte durch:

$$(x_A, x_B) \rightarrow (x_A^2, \sqrt{2 \cdot x_A \cdot x_B}, x_B^2)$$

Formel 5.3: Transformation durch diesen Kernel

– der Featureraum hat sich um eine Dimension erhöht und ist nun linear trennbar. Welche Funktionen als Kernel zulässig sind und wie genau die Hyperebene berechnet wird, ist in [BUR] sehr anschaulich beschrieben.

Wenn man mehrere Klassen mit SVMs unterscheiden will, kann man entweder den »1 versus 1«- oder den »1 versus Rest«-Ansatz verfolgen. Beim ersten wird eine SVM für den Vergleich jeder Klasse mit jeder anderen verwendet, beim zweiten eine für den Vergleich jeder Klasse mit dem kompletten Rest. Die erste Methode ist rechenintensiver (Komplexität quadratisch zur Anzahl der Klassen), die zweite liefert schlechtere Ergebnisse [KDD]. In Weka ist für SVMs der SMO-Algorithmus (Sequential Minimal Optimization) implementiert, der Klassen paarweise vergleicht.

Bewertung der Algorithmen

Joachims [JOA, JOA2] hat in mehreren Versuchen gezeigt, das SVMs die am besten geeignete Methode zur thematischen Textklassifizierung sind. In einem dieser Versuche wurden Texte aus dem Reuters-Korpus mit unterschiedlichen Verfahren in zehn Themengebiete klassifiziert und die Ergebnisse verglichen. Getestet wurden unter anderem Naive Bayes, k-NN, Entscheidungsbaum und SVMs, wobei letztere am besten abschneiden [JOA2]. Die Gründe dafür bestehen in der Tatsache, dass viele Features (nämlich die Häufigkeit der vorkommenden Wörter) vorhanden sind, wobei allerdings in jeden Dokument nur für wenige ein Wert vorhanden ist. SVMs sind unempfindlich gegen solche hochdimensionalen und dünn besetzen Feature-Vektoren. Dewdney et al. konnte dieses Ergebnis auch für die Genre-Klassifikation bestätigen [DEW].

Das schlechte Ergebnis des Naive-Bayes-Klassifikators erklärt sich durch die Annahme einer statistischen Unabhängigkeit der einzelnen Features, was bei Wörtern in Texten aber nicht der Fall ist [DEW, JOA]: Die Anzahl der Kommas hängt von der Satzlänge ab, der Anteil der Adjektive in einem Text steigt mit der Anzahl der Verben etc. Dieses Problem kann man dadurch verringen, dass man

dem Klassifikator geeignete Kombinationen von Features zur Verfügung stellt, beispielsweise die Anzahl der Adjektive relativ zu der der Verben.

5.2 Eigene Klassifikation

Wie in den vorigen Kapiteln beschrieben, gibt es für jedes Genre einen eigenen Klassifikator, der anhand verschiedener Features erkennt, ob ein Text zum jeweiligen Genre gehört oder nicht. Bei diesem Verfahren kann es passieren, dass ein Text mehreren Genres zugeordnet wird. Im Allgemeinen ist die Beschränkung auf eine Klasse auch nicht sinnvoll, da einige Dateien tatsächlich zu mehreren Genres gehören. In manchen Anwendungen kann dies wiederum eventuell erforderlich sein.

Um von den Abhängigkeiten der Genres voneinander zu profitieren, kann man die Einzelklassifikatoren miteinander kombinieren und so ausschließen, dass Texte einer bestimmten Klasse fälschlicherweise als eine andere erkannt werden. Die drei verwendeten Verfahren werden im folgenden vorgestellt.

Filtern

Um falsch klassifizierte Texte eines bestimmten Genres gezielt aus einer anderen Klasse zu entfernen, können Filter verwendet werden. Diese verbessern damit die Precision eines Einzelklassifikators. Die Regeln sind von der Form »Wenn der Text als A erkannt wurde, dann ist er nicht B«, oder an einem konkreten Beispiel: »Ist der Text ein Interview, so ist er kein Drehbuch.« Sie sind besonders wirkungsvoll, wenn viele A-Texte als B erkannt werden und umgekehrt wenige B als A. Hat der Klassifikator von A nur eine geringe Precision – sind also viele Texte aus anderen Genres darin eingeordnet – so werden möglicherweise korrekt klassifizierte Dateien aus B entfernt. Wenn im obigen Beispiel viele Drehbücher als Interview erkannt werden, so werden diese alle aus der Drehbuch-Klasse gelöscht. Statt die Precision zu erhöhen, verringert sich in diesem Fall der Recall.

Beim Bestimmen der Filterreglen ist also erstens darauf zu achten, dass die Klasse A eine hohe Precision hat und zweitens viele daraus fälschlicherweise als B erkannt werden. Als drittes Kriterium sollte noch beachtet werden, ob die Regel auch sinnvoll ist. Alle Code-Listings aus der Forum-Klasse zu löschen wäre beispielsweise nicht gut, da es viele Texte gibt, die beides sind.

…dann nicht	Regeln	rel. zu Genres
A	86	10,8
B	24	8,0
C	67	7,4
D	20	6,7
E	27	6,8
F	26	6,5
G	3	3,0
Gesamt	253	7,9

Tabelle 5.1: Regeln pro Hauptgruppe absolut und relativ zur Anzahl der klassifizierten Genres der Gruppe

Um Regeln zu finden, die diesen Anforderungen genügen, kann man aus den Ergebnissen des Trainings eine Konfusionsmatrix erstellen, um zu erkennen, welche Texte falsch in welchen Klassen landen. Alle, die nur in eine Richtung fehlerhaft klassifiziert werden, eignen sich als Filter. Im An-

hang ist eine Liste der verwendeten Filter angefügt. Die folgende Tabelle zeigt eine Zusammenfassung, wie viele Regeln es pro Hauptgruppe gibt.

Der Vorteil dieser Methode ist, dass die Abhängigkeiten zwischen einzelnen Klassen berücksichtigt werden, statt allgemeine Eigenschaften wie F1-Wert (siehe unten). Außerdem steigt die Qualität der Gesamtklassifikation beim Optimieren jedes einzelnen Klassifikators an. Es wird nicht nur diese eine Klasse verbessert, sondern jede davon durch Filterreglen abhängige. Gegebenenfalls können sogar neue Regeln hinzugefügt werden. Im Endergebnis können Dateien immer noch in mehreren Klassen vorkommen.

Reihenfolge, bestimmt durch Abhängigkeiten und Recallwerte

Statt erst alle Klassifikationen zu berechnen und anschließend die Ergebnisse zu filtern, kann man auch eine optimale Auswertungsreihenfolge bestimmen. Wird ein Text erkannt, so stoppt der Prozess. Dies hat zur Folge, dass keine Mehrfachklassifikationen entstehen.

Als erstes werden diejenigen Klassifikatoren angewendet, die besonders hohe Recall- und Precision-Werte haben und die folglich nur wenige Dateien fälschlicherweise erkennen. Es folgen die anderen Klassifikatoren, wobei Abhängigkeiten berücksichtigt werden: Erkennt ein Programm für

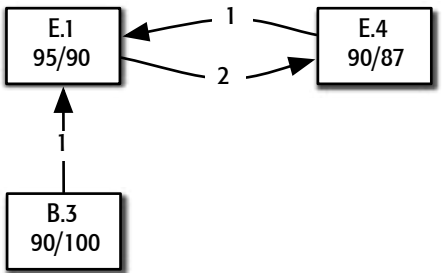

Abbildung 5.3: Ausschnitt aus dem Abhängigkeitsgraph

Klasse A auch Texte aus Klasse B, so ist es dahinter einzuordnen. Eine unvorteilhafte Reihenfolge führt zu geringer Präzision bei den vorderen Klassen (viele falsche Dateien werden erkannt) und geringem Recall bei den anschließenden, da diese Dateien dort gar nicht mehr ankommen.

Um die Reihenfolge zu bestimmen, wurde ein Abhängigkeits-Graph gezeichnet. Jede Klasse ist ein Knoten, der mit seinen Recall- und Precision-Werten versehen wird. Werden Texte aus einer Klasse (A) in einer anderen (B) erkannt, führt das zu einer gerichteten Kante von A nach B, die mit der Anzahl der Texte gelabelt wird. Theoretisch muss dieser Graph nur anhand dieser Kanten durchlaufen werden, um die optimale Reihenfolge zu erhalten. Leider enthält der Graph Zyklen, die aufgelöst werden müssen (vgl. Abb. 5.3). Dabei werden die Kanten mit kleineren Werten zuerst verfolgt, da dabei weniger Texte in eine falsche Klasse einsortiert werden. Sind die beide Werte in beide Richtungen gleich groß, wird zuerst der Knoten mit dem höheren Recall gewählt, da auf diese Weise mehr Texte nicht mehr in einer inkorrekten Kategorie landen. Auf diese Art entstand folgende Reihenfolge:

G ▸ E.2 ▸ F.4 ▸ F.2 ▸ F.3 ▸ C.9 ▸ C.6 ▸ C.5 ▸ B.3 ▸ B.1 ▸ D.1 ▸ D.3 ▸ D.2 ▸ E.4 ▸ E.1 ▸ E.3 ▸ C.8 ▸ C.1 ▸ A.5 ▸ A.7 ▸ A.8 ▸ C.4 ▸ C.3 ▸ C.2 ▸ A.3 ▸ A.4 ▸ A.1 ▸ B.2 ▸ A.2 ▸ C.7 ▸ A.6 ▸ F.1

Abbildung 5.4: Optimale Auswertungsreihenfolge

Eine andere Herangehensweise wäre, automatisiert jede Reihenfolge zu testen und diejenige zu wählen, die die besten Ergebnisse liefert.

Auswahl nach höchsten F1-Wert

Eine weitere Methode zur Bestimmung der besten Klasse, besteht darin, diejenige zu wählen, deren Klassifikator im Training den höchsten F1-Wert erreicht. Dem liegt die Annahme zugrunde, dass dieser mit der größten Sicherheit die richtige Klasse für eine Datei erkennt. Für die Berechnung des F1-Werts werden die Precisionwerte für die Originalklassen verwendet. Wie oben ergibt sich eine Reihenfolge, in der die Klassifikatoren durchlaufen werden können:

G ▸ F.2 ▸ C.9 ▸ E.2 ▸ F.4 ▸ D.3 ▸ B.3 ▸ B.1 ▸ C.5 ▸ D.1 ▸ A.5 ▸ C.8 ▸ F.3 ▸ E.1 ▸ A.7 ▸ D.2 ▸ A.3
▸ C.4 ▸ E.3 ▸ C.1 ▸ A.6 ▸ C.3 ▸ E.4 ▸ B.2 ▸ A.8 ▸ A.2 ▸ C.7 ▸ C.6 ▸ A.1 ▸ A.4 ▸ C.2 ▸ F.1

Abbildung 5.5: Rangfolge der F1-Werte

6 Evaluierung

Zur Bewertung von Klassifikatoren existieren mehrere Gütemaße, die hier kurz vorgestellt werden. Weiterhin wird geklärt, auf welche Art und Weise Sonderfälle behandelt werden sollen. Da die Klassifikation durch eine Reihe von Einzelprogrammen erfolgt, wovon jedes überprüft, ob eine Datei in »seine« Klasse gehört, werden manche Dateien teils in mehrere Gruppen eingeordnet, teils überhaupt nicht erkannt. Dadurch ergeben sich drei unterschiedliche Berechnungsbasen: die Gesamtanzahl der Dateien, die der gefundenen Dateien und die Zahl der Einordnungen.

Außerdem besteht noch die Möglichkeit, dass eine Datei zwar nicht ursprünglich einer bestimmten Klasse zugeordnet wurde, dort jedoch trotzdem richtig ist. Ein Beispiel hierfür ist eine Personenliste mit statistischen Informationen. Dadurch ergeben sich auch noch verschiedene Arten von »richtig«: die korrekte Erkennung in der Originalklasse und die Erkennung als ein Genre, das auch stimmt.

Accuracy und Classification Error

Die *Klassifikationsgenauigkeit* (*Accuracy*) gibt an, wie viele Objekte der Testmenge der richtigen Klasse zugeordnet wurden; der *Klassifikationsfehler* (*Classification Error*) wie viele der falschen.

Accuracy = richtig erkannte Dateien aus Klasse A / Gesamtzahl Dateien aus A

Class. Error = falsch erkannte Dateien aus A / Gesamtzahl Dateien aus A

Formel 6.1: Formeln für Accuracy und Classification Error [KDD, S. 110]

Am aussagekräftigsten erscheint mir hier die Berücksichtigung *aller* richtigen Klassifikationen und eine Berechnung auf Basis der *Einordnungen*. Um meine Ergebnisse auch mit denen anderer vergleichen zu können werden alle Werte zusätzlich nur mit den Texten der Originalklassen berechnet. Die nicht gefundenen Dateien werden nicht berücksichtigt. Bei mehrfach erkannten Files zählt jede richtige bzw. falsche Klasse, da sonst Dateien die einmal richtig und einmal falsch sind doppelt gewertet würden, alle anderen nur einmal.

Recall, Precision und F1-Metrik

Der *Recall* gibt an, wie viele Objekte einer Klasse korrekt zugeordnet wurden, *Precision* bezeichnet den Anteil der richtigen Objekte in einer Klasse. Diese beiden Werte lassen sich kombinieren, beispielsweise in der *F1-Metrik*:

Recall = richtig erkannte Dateien aus Klasse A / Gesamtzahl Dateien aus A

Precision = Richtig erkannte Dateien aus A / alle als A erkannte Dateien

F1 = 2 · (Precision · Recall) / (Precision + Recall)

Formel 6.2: Formeln für Recall, Precision und F1 [DEW]

Beim Recall werden nur die Texte der Originalklasse gewertet, als Basis dient jeweils die Dateianzahl des betrachteten Bereichs. Für die Precision werden – jeweils auf Basis der Einordnungen – zwei Werte berechnet: für die Dateien der Originalklasse und für alle korrekt erkannten Texte. Grundlage für die Berechnung der F1-Metrik ist der letztgenannte Wert.

6.1 Eigene Klassifikatoren

Bei der folgenden Auswertung werden die Klassen Zitate (D.4), Wortliste (E.5) und sontige Listen (E.6) nicht berücksichtigt und auch von der Gesamtzahl der Dateien abgezogen. Es gibt also insgesamt 32 Klassen mit 630 Dateien.

6.1.1 Gesamtergebnis

Insgesamt werden 74,8% (471) der Dateien erkannt, der Rest konnte keiner Klasse zugeordnet werden. Einige der Dateien werden mehrfach erkannt, wodurch sich eine Gesamtzahl der Einordnungen von 694 ergibt.

Genre	erk.	eing.	Genre	erk.	eing	Genre	erk.	eing.	Genre	erk.	eing.
A.1	14	16	**B.1**	15	25	**C.6**	8	12	**E.2**	18	24
A.2	15	19	**B.2**	16	20	**C.7**	10	14	**E.3**	15	16
A.3	15	18	**B.3**	12	16	**C.8**	15	22	**E.4**	14	24
A.4	10	13	**C.1**	16	35	**C.9**	18	30	**F.1**	11	17
A.5	17	30	**C.2**	16	18	**D.1**	16	17	**F.2**	16	25
A.6	11	14	**C.3**	15	18	**D.2**	16	22	**F.3**	17	41
A.7	13	29	**C.4**	15	24	**D.3**	13	16	**F.4**	19	28
A.8	15	23	**C.5**	17	25	**E.1**	13	22	**G**	20	21

Tabelle 6.1: Anzahl der erkannten Dateien (erk.) und Einordnungen (eing.) je Genre

Accuracy und Classification Error

Etwa zwei Drittel (454 bzw. 65,4%) der Zuordnungen sind richtig, die Mehrzahl (381) davon in der Originalklasse. Der Anteil der korrekt klassifizierten Dateien variiert stark zwischen den einzelnen Genre-Gruppen und reicht von 53,1% für Journalismus bis 80,2% für Verzeichnisse (und 95,2% für »Nichts«).

Durch simultanes Anwenden der in der Trainingsphase ermittelten Filterregeln kann das Ergebnis auf 75,2% der erkannten Dateien verbessert werden. Am meisten profitiert hierbei der Bereich Kommunikation, in dem jetzt 74,3% statt davor 58,6% richtig erkannt werden. Der Grund dafür ist, dass 22 falsche Dateien aus der Brief/Rede-Klasse entfernt werden. Es gehen zwar auch korrekt klassifizierte Dateien verloren; deren Zahl ist mit 33 für den gesamten Korpus (davon 13 in den Originalklassen) allerdings vergleichsweise gering.

Wenn eine Einordnung in eine einzige Klasse erzwungen werden soll, kann dazu wie oben (5.2) geschildert der F1-Wert der Trainingsklassifikation oder die durch Abhängigkeiten bestimmte Reihenfolge gewählt werden. Hier werden 78,6% bzw. 80,5% der eingeordneten Dateien richtig erkannt. Die Verbesserung gegenüber der Filter-Methode liegt hauptsächlich darin begründet, dass die Gesamtzahl der erkannten Dateien kleiner ist. Absolut gesehen verringert sich die Zahl der korrekt identifizierten Texte von 422 auf 370 bzw. 379. Berücksichtigt man nur die Originalklassen, dann sinkt sie von 368 auf 327 bzw. 340.

Die folgende Tabelle zeigt die Werte für Accuracy (jeweils für Originalklasse und alle richtig klassifizierten Dateien) und Classification Error. Als Basis dienen alle erkannten Dateien einer Gruppe (zum Beispiel alle Erkannten aus Gruppe A, egal als was sie klassifiziert wurden).

Vergleich mit dem Zufall

Bei 32 Klassen ist die Wahrscheinlichkeit der Einordnung in die richtige Klasse 1/32 = 3,1%. Da hier jeder Text nur in eine Klasse eingeordnet wird, ist ein Vergleich mit den beiden Mehrfachklassifikationen nicht möglich. Außerdem werden nur die Dateien, die in ihrer Original-Klasse erkannt werden, berücksichtigt; die Basis sind hier ausnahmsweise alle 630 Dateien. Der Vergleich mit der Auswahl der Klasse nach F1-Wert oder Auswertungsreihenfolge zeigt, dass die Verfahren mit 51,9% bzw. 54,0% etwa 17 mal besser als der Zufall sind.

	Mehrfachklassifikation			mit Filter			Reihenfolge			F1-Auswahl		
	Acc. (Org.)	Acc. (alle)	Error	Acc. (Org.)	Acc. (alle)	Error	Acc. (Org.)	Acc. (alle)	Error	Acc. (Org.)	Acc. (alle)	Error
Gesamt	54,9	65,4	34,6	65,6	75,2	24,8	72,2	80,5	19,5	69,4	78,6	21,4
A. Journalismus	41,4	53,1	46,9	49,2	62,1	37,9	56,0	69,7	30,3	57,3	69,1	30,9
B. Literatur	60,7	65,6	34,4	76,6	83,0	17,0	74,4	81,4	18,6	74,4	81,4	18,6
C. Information	57,1	68,7	31,3	68,1	79,4	20,6	71,8	80,9	19,1	67,7	77,7	22,3
D. Dokumentation	58,2	69,1	30,9	61,5	69,2	30,8	68,9	75,6	24,4	66,7	75,6	24,4
E. Verzeichnis	66,3	80,2	19,8	71,1	82,9	17,1	85,0	90,0	10,0	75,0	85,0	15,0
F. Kommunikation	49,5	58,6	41,4	70,3	74,3	25,7	81,0	85,7	14,3	77,8	84,1	15,9
G. Nichts	95,2	95,2	4,8	100,0	100,0	0,0	100,0	100,0	0,0	100,0	100,0	0,0

Tabelle 6.2: Mittelwerte für Accuracy und Classification Error (in %)

Recall, Precision und F1

Wie erwartet schneidet die Mehrfachklassifikation mit Filter am besten ab, sie erreicht einen F1-Wert von 65,6%, wobei der Recall mit 58,4% nur knapp unter der besten Wert (60,5%, ungefiltert) liegt und die Precision mit 74,9% ebenfalls gut ist. Die anderen Verfahren sind nicht wesentlich schlechter, der niedrigste F1-Wert ist 62,5%.

	Mehrfachklassifikation			mit Filter			Reihenfolge			F1-Auswahl		
	Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1
Gesamt	60,5	65,4	62,9	58,4	74,9	65,6	54,0	80,5	64,6	51,9	78,6	62,5
A. Journalismus	41,9	50,3	45,7	40,6	59,7	48,3	38,1	71,0	49,6	39,4	69,6	50,3
B. Literatur	61,7	66,1	63,8	60	78,3	67,9	53,3	78,0	63,3	53,3	76,2	62,7
C. Information	66,5	71,0	68,3	64,1	79,2	70,9	55,3	80,3	65,5	51,8	77,8	62,2
D. Dokumentation	53,3	80,0	64,0	53,3	81,4	64,4	51,7	85,0	64,3	50,0	80,5	61,7
E. Directory	71,3	77,3	74,2	67,5	84,8	75,2	63,8	91,0	75,0	56,3	88,3	68,8
F. Kommunikation	57,5	57,9	57,7	65,0	71,6	68,1	63,8	78,3	70,3	61,3	80,0	69,4
G. Nichts	100	95,2	97,5	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Tabelle 6.3: Mittelwerte für Recall, Precision und F1 (in %)

Betrachtet man die einzelnen Gruppen, so stellt man fest, dass nicht nur die Werte stark schwanken, sondern auch das beste Verfahren variiert: Beim Journalismus erweist sich das F1-Auswahlverfahren als das beste, bei Kommunkation wiederum die Berücksichtigung von Abhängigkeiten und ansonsten die gefilterte Mehrfachklassifikation. Das schlechte Abschneiden des Filterns beim Journalismus ist überraschend, da es hier pro Genre relativ viele Regeln gibt, nämlich 10 statt der

durchschnittlichen 6 (vgl. Tabelle 5.1). Trotzdem werden nicht etwa zu viele richtige Dateien ausgefiltert – der Recall liegt sogar über den beiden anderen Verfahren – sondern vielmehr ist die Precision gering. Auch die Performance der zum Filtern herangezogenen Einzelklassifikatoren ist nicht schlechter als bei den anderen Gruppen. Vermutlich sind die Filterregeln nicht optimal. Auch bei den Kommunkationstexten sind die Regeln nicht ausreichend. So gibt es beispielsweise keine Regel, die falsche Dateien aus der Forum-Klasse filtert, da beim Training kein Bedarf dafür zu erkennen war. Zusätzlich wurden die Klassen, die zum Filtern für Briefe verwendet werden selbst nur schlecht erkannt (nur zwei Glossen wurden gefunden), weswegen die Regeln nicht wirksam wurden.

Die Werte für Recall, Precision und Accuracy für die einzelnen Genres befinden sich im Anhang.

6.1.2 Einzelklassifikatoren

Die Güte der einzelnen Klassifikatoren ist sehr unterschiedlich und reicht von einem F1-Wert von 14,7% für Glossen bis 100,0% für »Nichts«. Dies liegt daran, dass die Texte tatsächlich unterschiedlich schwer zu erkennen sind. Glossen haben keine feste äußere Form, sind durchschnittlich lang, weisen keine spezifische Sprache (weder im Vokabular noch in Zeit, lediglich begrenzt im Stil) und auch keine HTML-Besonderheiten auf. Wie in 4.2 schon erwähnt, werden die Genres, welche eine feste Form besitzen (Verzeichnisse, Gedichte, FAQs, Foren), überdurchschnittlich gut erkannt.

Durch Analyse der Konfusionsmatrix (siehe Abb. 9.1 im Anhang) kann man feststellen, wo die Schwächen der Klassifikatoren liegen und wie schwerwiegend die Fehler sind, die gemacht werden. Zu den harmloseren Fehlklassifikationen gehören das Erkennen von Feature als Glosse (4 Fehler) oder Kommentar (6), da beide zu den journalistischen Genres gehören, die die Meinung des Autors zum Ausdruck bringen können. Auch die Klassifikation als Präsentation (4) und Erklärung (3) ist kein schwerwiegender Fehler, da im Feature ebenfalls eher subjektive Informationen zu einem bestimmten Thema dargestellt oder Sachverhalte erklärt werden. Teilweise entscheidet einzig die Aktualität des beschriebenen Geschehens darüber, ob etwas als Feature oder als Beschreibung historischer Ereignisse gewertet werden soll. Ähnlich verhält es sich mit der Fehlklassifikation von wissenschaftlichen Texten als Feature (4), da diese auch »Analysen und Hintergründe« (vgl. 2.2.2) liefern – nur eben zu wissenschaftlichen statt zu gesellschafltichen oder politischen Themen. Auch die Klassifikation von Kommentaren als Nachrichten ist akzeptabel, da erstere schließlich auch auf aktuelle Themen Bezug nehmen. Drei Erzählungen/Romane wurden als Porträt identifiziert, einer als Reportage. Dies lässt sich damit erklären, dass auch Romane das Leben von Personen oder bestimmte Erlebnisse schildern, allerdings von fiktiven. Eine Verwechslung ist hier verzeihlich. Weitere häufig auftretende aber vertretbare Fehler sind die Einordnung von offiziellen Berichten in das Präsentations-Genre (oft enthalten sie eine Darstellung der Firma oder Organisation) und die von Blogs in die Glossen-Klasse (relativ häufig werden in Blogs auch in gewisser Weise Glossen geschrieben).

Bei einigen Genre-Paaren findet man besonders häufig korrekte Doppelklassifikationen, zum Beispiel Personen und Statistik (4 bzw. 2 in anderer Klasse) oder Foren und Code (4 Code-Listings sind auch Forum). Tabelle 6.4 stellt einige Fehlklassifikationen dar, die beseitigt werden sollten.

Wie erwartet sind die Klassifikatoren, die von WEKA bestimmte Merkmale enthalten – Erklärung (Recall 35,0%, Precision 57,1%) und Rezension (40,0% und 72,2%) – relativ schlecht. Dies bestätigt die

These, dass, zumindest mit dem vorhandenen kleinen Korpus, beliebige statistische Features zu keinen guten Ergebnissen führen.

ist	erkannt als	Anzahl	Bemerkung
A.5	B.3	2	ähnliche Sruktur
A.4	F.1	4	persönliche Ansprache, viele »I« und »you«
A.5	A.4	4	
A.5	F.1	5	enthält Begrüßung und Abschied, Verbesserung durch Zählen der Interviewpartner
B.1	F.1	5	
B.3	A.5	1	ähnliche Struktur
C.1	A.5	4	ein wissenschaftlicher Text enthält Glossar
C.9	C.6	4	Codewörter und Variablennamen als Fremdwörter erkannt
D.1	C.2	4	
F.2	E.4	5	enthält aufeinanderfolgende Datumsangaben
F.3	B.2	4	Blog könnte Erzählungen enthalten, ist aber hier nicht der Fall
F.3	E.4	4	enthält aufeinanderfolgende Datumsangaben
F.3	F.1	8	persönliche Ansprache, viele »I« und »you«

Tabelle 6.4: Häufige Fehler in der Klassifikation

6.1.3 Filterreglen

Filterregeln sind dann nützlich, wenn möglichst viele falsche und möglichst wenig richtige Texte aus den einzelnen Klassen entfernt werden (vgl. 5.2). Ihre Effektivität wächst außerdem mit dem Recall der Klasse, die im Kopf der Regel steht (in »wenn A, dann nicht B« wäre dies Klasse A). Aus diesen zwei Vorgaben lässt sich ein Maß für die Güte des obigen Filters bilden:

$$\text{Güte} = \text{Recall von A} \cdot (\text{falsch aus B entfernte Dateien} - \text{richtig aus B entfernte Dateien})$$

Formel 6.1: Maß für die Güte von Filterregeln

Dieses Maß kann beliebige positive und negative Werte annehmen; Je größer er ist, desto besser ist die Regel. Ist der Wert negativ, so werden mehr richtige als falsche Dokumente entfernt, der Filter verschlechtert das Ergebnis. Die Analyse einiger Regeln lieferte recht gute Ergebnisse: Der beste Werte wird beim Filtern von Blogs aus Briefen (+5,6) erzielt, der schlechteste beim Entfernen von Foren aus Blogs (-1,6). Bildet man den Mittelwert aus den Gütemaße der besten und schlechtesten zehn (vgl. Tabelle 6.5), erhält man +2,2 – ein zufriedenstellendes Ergebnis.

wenn	dann nicht	Güte	wenn	dann nicht	Güte
F.3	F.1	5,6	F.2	F.3	-1,6
A.5	F.1	3,5	A.3	A.4	-1,1
B.1	F.1	3,5	E.3	D.1	-0,75
A.5	A.4	2,8	A.5	A.2	-0,7
F.3	B.2	2,8	A.5	C.4	-0,7
F.3	E.4	2,8	C.4	A.1	-0,7
F.2	E.4	2,4	C.4	C.7	-0,7
A.7	A.1	2,1	A.7	A.6	-0,7
C.1	A.7	2,1	A.7	F.1	-0,35
F.3	A.4	2,1	A.4	A.2	-0,2

Tabelle 6.5:
Die 10 besten und schlechtesten Filter

Anhand der Konfusionsmatrix kann man erkennen, dass auch einige Filter fehlen. Zum einen liegt dies daran, dass beim Erstellen der Regeln deren Notwendigkeit nicht erkannt wurde, zum anderen, dass im Trainkorpus Fehlklassifikationen in der Gegenrichtung des Tests vorkamen. So wurden im Training Foren als Blogs erkannt, später war genau das Gegenteil der Fall. Ein größerer Trainkorpus und eine systematischere Zusamenstellung der Filter sollten diese Mängel beseitigen können.

6.1.4 Klassifikation in Hauptgruppen

Fasst man die Texte einer Gruppe zu einer einzigen Klasse zusammen, so kann man erkennen, wie gut die Klassifikatoren der Subgenres geeignet sind, die Hauptkategorie zu erkennen. Es werden alle Texte als richtig gewertet, die in der korrekten Gruppe erkannt wurden. Die folgende Tabelle (6.5) zeigt Recall, Precision und Accuracy für diese Klassifizierung.

Für Journalismus und Information/Wissen ergibt sich die stärkste Verbesserung. Dies liegt daran, dass diese die meisten Dateien beinhalten. Wenn man annimmt, dass alle 42 jetzt beseitigten Fehler gleichmäßig auf die 32 Genres verteilt werden, so landen dort 14 bzw. 15. Ein tatsächlicher Anstieg der richtig erkannten Dokumente (zusätzliche 17!) findet nur beim Journalismus statt. Betrachtet man die Konfusions-Matrix (Tab. 6.6), so stellt man fest, das tatsächlich viele Dateien innerhalb diese Bereichs fehlklassifiziert werden.

	Recall	Precision		Accuracy		F1
		Original	alle	Original	alle	alle
Gesamt	65,1	77,8	84,1	77,8	84,1	73,4
A. Journalismus	57,5	75,4	79,5	78,0	81,4	66,7
B. Literatur	61,7	80,4	80,4	78,7	85,1	69,8
C. Information	68,8	79,6	86,4	77,5	86,8	76,6
D. Dokumentation	55,0	76,7	83,7	64,7	72,5	66,4
E. Verzeichnis	70,0	78,9	91,5	78,9	85,9	79,3
F. Kommunikation	68,8	70,5	78,2	79,7	84,1	73,2
G. Nichts	100,0	100,0	100,0	100,0	100,0	100,0

Tabelle 6.5: Recall, Precision und Accuracy für die Klassifikation mit Filterung in Hauptgruppen

Die restlichen Fehler erschließen sich aus den Problemen der Einzelklassifikatoren: Roman und Erzählung werden als Feature, Reportage oder Porträt erkannt, Features als Präsentation oder Erklärung, und Forum und Blog als Timeline. Von Journalismus nach Kommunikation und umgekehrt werden die meisten Dateien falsch klassifiziert. Dies liegt hauptsächlich an dem suboptimalen Brief/Rede-Erkenner, der alleine für 19 dieser Fehler verantwortlich ist.

	A	B	C	D	E	F	G
A	104	12	21	7	4	14	1
B	3	40	5			6	
C	18		139	12	9	6	
D	2		9	33	1		
E	3	1	8	2	66	14	
F	13	8	10		6	70	
G							20

Abbildung 6.1: Konfusionsmatrix für Mehrfachklassifikation in Hauptgenres (Spalten: ist Klasse; Reihen: erkannt als)

Insgesamt finden relativ wenige Fehlklassifikationen in andere Hauptgruppen statt, woraus man schließen kann, dass die hierarchische Gliederung der Genres durchaus effizient ist.

6.2 Automatische Klassifikatoren

Eine Alternative zu obigen Verfahren ist die Verwendung der in 5.1 beschriebenen Algorithmen. Dazu wurden für alle Dateien sämtliche in den Einzelklassifikatoren vorkommenden Features berechnet, wobei die Texte, für die es keinen Erkenner gibt weggelassen wurden (also die aus D.4, E.5 und E.6). Die Gesamtzahl der Dateien beträgt deshalb wieder 630. Eine Aufzählung der knapp 200 verwendeten Merkmale befindet sich im Anhang. Mit Hilfe von Weka wurden die Klassifikatoren anhand der Trainingsdaten erstellt und anschließend ihre Qualität mit den Testdaten überprüft.

Am besten schneiden Support Vector Machines ab: Mit einem linearen Kernel werden 48,3% der Texte richtig eingeordnet. Es folgen Naive Bayes (45,4%) und J48-Entscheidungsbaum (37,9%), das Schlusslicht bildet mit einer Erkennungsrate von nur 32,2% der Nearest-Neighbour-Klassifikator. Auffallend ist, dass die verschiedenen Verfahren unterschiedlich gut für die einzelnen Genres funktionieren (vgl. Tabelle 9.10 im Anhang). Bei zweisprachigen Wörterbüchern, wo k-NN keinen einzigen Text richtig erkennt, erzielen SVMs immerhin F1-Werte von 33,3%; das eher schlechte Ergebnis der Support Vector Machines bei Feature und offiziellem Bericht wird von Naive Bayes deutlich übertroffen. Daraus kann man folgern, dass eine Kombination der verschiedenen Klassifikatoren zu deutlich besseren Ergebnissen führt. In [SCU] sind mehrere Möglichkeiten dafür beschrieben. Im einfachsten Fall wählt man für einen Text diejenige Klasse aus, für die die meisten Klassifikatoren sich entscheiden. Verfeinern lässt sich dies durch eine Gewichtung der einzelnen Verfahren. Hier könnte man beispielsweise allgemein das SVM-Ergebnis höher werten, da dieser insgesamt die besten Ergebnisse erzielte, oder einen Text in »Nichts« einordnen, falls der Bayes-Klassifikator dies vorschlägt, da dieser hier mit 85,7% sehr gute Werte erreicht. Dies umzusetzen und auszuwerten wäre allerdings überaus aufwändig und daher im Rahmen der Magisterarbeit nicht möglich.

Insgesamt ist das Ergebnis der automatischen Verfahren nicht so positiv wie erwartet. Erzwingt man bei meinem eigenen Verfahren eine Einordnung in nur eine Klasse (z.B. Auswahl nach Reihenfolge) und wählt als Basis wieder alle 630 Dateien, so werden immerhin noch 54,0% der Texte richtig erkannt – deutlich über dem Wert der Support-Vector-Machines. Allerdings wäre es möglich, dass deren Performance durch die Wahl anderer Kernels noch steigt. auf Grund fehlender Rechenleistung konnte hier bisher nur mit linearen Kernels gearbeitet werden.

Das schlechte Abschneiden der automatischen Klassifikations-Algorithmen wird auch dadurch bedingt, dass der Trainingskorpus zu klein ist. Joachims verwendet in seinem Experiment [JOA2] immerhin 9 603 Trainingsdokumente, also knapp 1 000 für jede Klasse. Außerdem würde [DEW] zufolge das Ergebnis besser ausfallen, wenn keine Einordnung in eine Klasse erzwungen wird.

Besonders das schlechte Abschneiden des Nearest-Neighbour-Klassifikators überrascht, da dies laut Joachims [JOA2] nach SVMs die beste Methode zur Textklassifikation ist.

Analyse des Entscheidungsbaums

Der J48-Algorithmus erzeugt einen Entscheidungsbaum mit 132 Blättern und 263 Knoten. Im abgebildeten Ausschnitt aus dem Baum (Abb. 6.2) kann man erkennen, dass relativ viele Texte aus ei-

nem Genre schon nach wenigen Schritten klassifiziert werden. Die Auswahlmerkmale erscheinen, wie im folgenden für drei Genres exemplarisch gezeigt, durchaus sinnvoll und werden zum Teil auch in meinen Einzelklassifikatoren verwendet.

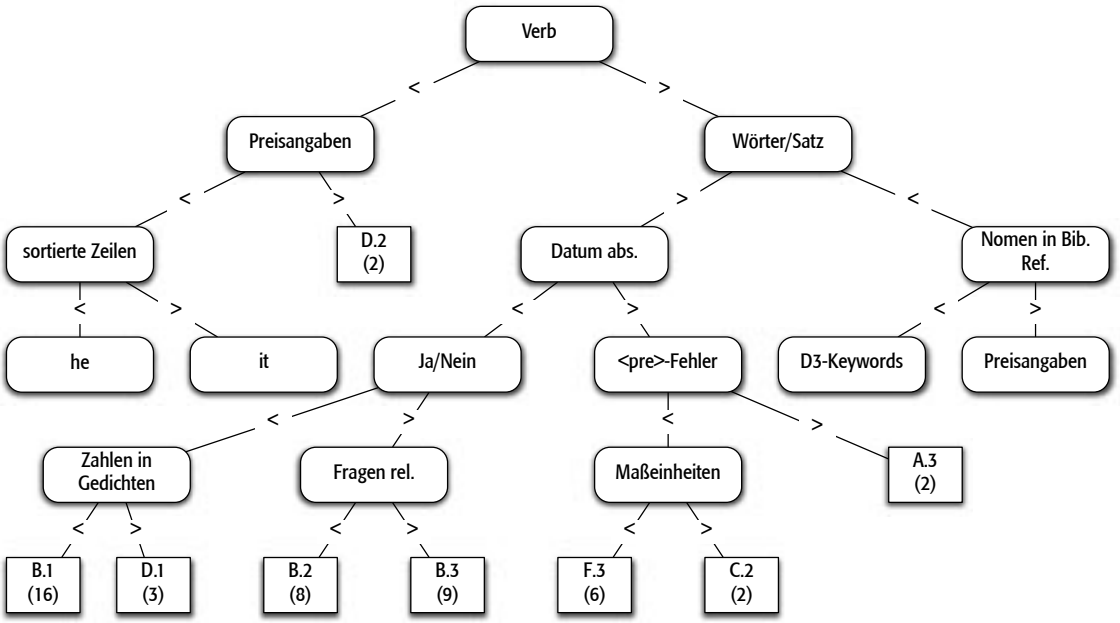

Abbildung 6.2: Die ersten Stufen des mit J48 erzeugten Baums

Genau wie dort enthalten *Kataloge* viele Preisangaben, zusätzlich wird eine geringe Anzahl an Verben verlangt. Das Erkennen von *Gedichten* erfolgt anhand zweier expliziter Gedichts-Features (vgl. 4.2.2), neue Kriterien sind wenige Datumsangaben und zustimmende/ablehnende Wörter sowie viele Verben – alles passende Merkmale. Allerdings wird das augenfälligste Feature, nämlich die Anzahl von Gedichtzeilen, nicht berücksichtigt. Bei den Auswahlkriterien für *Personenlisten* erstaunt es, dass die Anzahl der Namen für den Großteil der erkannten Texte nicht verwendet wird und statt dessen die Anzahl der »it« auschlaggebend ist. Trotz dieser Ungereimtheiten erreichen die drei Textklassen bei der Erkennung durch den Entscheidungsbaum gute F1-Werte von 50,0% bis 74,4%.

6.3 Vergleich mit anderen Ergebnissen

Andere Autoren erzielten mit automatischen Verfahren weit bessere Ergebnisse als ich in meinen Versuchen. Stamatatos et. al. erreichen mit nur 20 Trainingstexten für jedes der vier Genres und der Linearen Diskriminanz-Analyse (LDA) eine Accuracy von über 97%, und das nur unter Verwendung von 30 Wörtern und 8 Satzzeichen [STA]. Ein kleiner Test mit all meinen Features für die Genres A.4 (ähnlich Editorial), A.6 (Nachricht), A.8 (Reportage) und – in Ermangelung von Leserbriefen – A.5 ergab einen Wert für Bayes-Klassifikation von 72,5%. Bei einer Verringerung der Features auf insgesamt 30 Vokabular-Merkmale (u.a. Pronomen, Artikel, Konditional) und Satzzeichen verschlechterte er sich sogar auf 63,8%. Vermutlich stammten die Texte bei [STA] immer von den selben Autoren, das Themengebiet variierte nicht so stark oder es gab andere Ähnlichkeiten, die bei der Klassifizierung hilfreich waren. Anders lässt sich dieses Ergebnis kaum erklären.

Die Ergebnisse anderer Arbeiten erscheinen realistischer: So erreichte [DEW] bei einer Korpusgröße von knapp 10 000 Texten und sieben Genres (Ads, Bulletin Board, FAQ, Message Board, Radio, Reuters, TV) mit SVMs F1-Werte von bis zu 89,1%. Sinkt die Anzahl der Trainingstexte, so nimmt die Güte der Klassifikation ab. [WAS] erzielen mit einem Bayes-Klassifikator mit POS-Features bei 425 Texten und neun Klassen ähnlich denen des Brown-Korpus mittlere Recall- und Precisionwerte von 57,8% bzw. 62,2% und sind dabei etwas schlechter als meine Klassifikation (60,5% bzw. 65,4%), obwohl sie über deutlich weniger Genres verfügen. An den Ergebnissen von Karlgren und Cutting [KAC] sieht man den Einfluss der Genrezahl auf die Güte der Klassifikation: Bei vier Kategorien (Presse, Non-Fiction, Fiction, Sonstiges) beträgt die Accuracy mit LDA 73%, untersucht man die 15 Brown-Kategorien nur noch 52%. Zum Vergleich: die Accuracy, die durch Kombination meiner Einzelklassifikatoren erreicht wird, beträgt 54,0% (Originalklassen/alle Dateien) – betrachtet man die Zahl der richtigen Einordnungen sogar 80,3%.

Berücksichtigt man also die relativ geringe Größe des Trainingskorpus und die hohe Anzahl der Klassen, so kann sich mein Ergebnis durchaus sehen lassen.

7 Fazit

In dieser Arbeit wurde eine neues hierarchisches System von Genres entworfen. Mit Hilfe eines eigens dafür zusammengestellten Korpus wurden für jedes Genre dessen identifizierende Merkmale bestimmt und daraus ein Klassifikator entwickelt. Deren Ergebnisse wurden auf verschiedene Arten kombiniert, um die Performance des kompletten Systems zu steigern oder eine eindeutige Einordnung zu erreichen. Dieses Vorgehen stellt ein neues Verfahren zur automatischen Erkennung des Genres von Texten dar. Vergleiche mit klassischen Algorithmen aus dem Data Mining zeigen, das es diesen zumindest ebenbürtig ist – zumindest für relativ kleine Korpusgrößen.

7.1 Verbesserungspotenzial

Obwohl die Ergebnisse also alles in allem meinen Erwartungen entsprechen, können noch einige Dinge verbessert werden. Zum einen müssen die vier fehlenden Klassifikatoren – Zitate, Wortlisten, sonstige Listen und Kombinationen – noch programmiert und gegebenenfalls die Korpora ergänzt werden. Eine Idee für das Finden von Wortlisten wäre, den Text in gleichgroße Abschnitte aufzuteilen und in jedem die Anzahl der neu vorkommenden Wörter zu zählen. Bei natürlichen Texten nimmt diese nach hinten hin ab, bei Wortlisten, die ja eine eher wahllose Zusammenstellung von Begriffen sind, bleibt sie konstant. Erste Vorversuche konnten dies auch bestätigen

Für die Erkennung von Kombinationen müssen zunächst die verschiedenen Textblöcke identifiziert werden, um anschließend einzeln klassifiziert werden zu können. Auch die Qualität der anderen Klassifikatoren könnte dadurch verbessert werden, da zum Beispiel Einleitungs- und Hinweistexte und insbesondere die Navigationselemente entfernt werden könnten. Weitere Maßnahmen, die die Erkenner noch optimieren könnten, sind das Vervollständigen der Wortlisten und eine Erweiterung des Trainingskorpus. Die mangelhaften Klassifikatoren, speziell derjenige für Glossen, müssen noch einmal komplett überarbeitet werden. Auch der verwendete Tagger sollte ausgetauscht werden, da dieser unter anderem in den Phrasen »which is closely related to«, »will be discussed«, »specialized hardware« Past-Tense Verben erkennt.

Um die bisher als überhaupt nichts klassifizierten Dateien auch in ein Genre einzuordnen, könnte man automatische Methoden wie Support Vector Machines verwenden. Eine andere Möglichkeit wäre die Anpassung der Einzel-Klassifikatoren. Dabei dürfen jedoch nicht einfach die Bedingungen weniger streng sein, da ansonsten auch viele falsche Texte erkannt werden. Vermutlich müssen neue Features oder -Kombinationen hinzugenommen werden.

Die Genauigkeit der Gesamtklassifikation kann durch eine Überarbeitung der Auswertungsreihenfolge und der Filterreglen gesteigert werden. Eine Verbesserung der automatischen Klassifikation kann man durch einen größeren Trainingskorpus oder geeignete Kombinationen der unterschiedlichen Verfahren erreichen.

Schließlich könnte auch das Genre-System noch verbessert werden, da einige Texte nicht darin eingeordnet werden konnten.

7.2 Ausblick

Eine Möglichkeit um das Problem der Listen- und Blockerkennung bei allein nach grafischen Gesichtspunkten statt Strukturinformationen programmierten HTML-Seiten (vgl. 4.2.3) zu lösen, wäre, den Code wie ein Browser zu rendern. Die einzelnen Textabschnitte könnten durch eine optische Analyse des so entstandenen Bildes identifiziert werden.

Interessant wäre es auch, die Art der Bilder zu untersuchen: sind es Infografiken, Fotos von Personen/Handlungen oder reine Zierelemente? Ansatzpunkte liefern hier die Dateigröße, die Größe des Bildes und seine Position im Text sowie das Dateiformat (GIF, JPEG oder PNG). Bilder am Seitenanfang sowie sehr kleine Bilder sind oft nur Dekoration, Bilder mitten im Text sind Werbung oder haben mit dem Thema des Textes zu tun (Infografik, Foto), GIFs mit großen Ausmaßen aber kleiner Dateigröße deuten auf Infografiken mit wenigen Farben hin, wie sie oft in wissenschaftlichen Texten vorkommen. Animierte GIFs findet man hauptsächlich auf unprofessionellen Seiten oder als bewegte Smilies in Foren. Komplexere Bilderkennungsalgorithmen können zusätzlich bestimmen, was auf dem Bild dargestellt ist.

Zusätzlich zu den bisher berechneten Features wäre die Anzahl der Rechtschreibfehler auch eine Untersuchung wert. Diese Information kann auch hilfreich sein, um bei weiteren linguistischen Aufgaben gezielt Genres mit vielen Fehlern zu filtern.

Außerdem könnten die Einzelklassifikatoren, statt eine einfache Ja-Nein-Entscheidung zu treffen, die Wahrscheinlichkeit angeben, mit der ein Text zum jeweiligen Genre gehört. Nach dem Maximum-Likelihood-Prinzip könnte die passendste Klasse gewählt werden.

Danke

Bedanken möchte ich mich bei Christoph Ringlstetter und Prof. Klaus Schulz für interessante Anregungen, Kritik und Betreuung der Magisterarbeit. Vielen Dank auch an Christoph Koch für journalistischs Expertenwissen, an Manuel Robl, der für die Evaluation der Genres die Texte von Hand klassifizierte, und an Thomas Glöckler fürs Korrekturlesen.

8 Literaturverzeichnis

[ARG] Argamon, Shlomo; Koppel, Moshe; Fine, Jonathan; Shimoni, Anat Rachel (2003). Gender, Genre, and Writing Style in Formal Written Texts.

[BIB] Biber, Douglas (1989). Variation across speech and writing, Cambridge University Press, Cambridge

[BUR] Burges, Christopher J.C. (1998): A Tutorial on Support Vector Machines for Pattern Recognition. Aus: Data Mining and Knowledge Discovery 2, 121-167.

[CRK] Crowston, Kevin; Kwasnik, Barbara H. (2004). A Framework for Creating a Facetted Classification for Genres: Addressing Issues of Multidimensionality, hicss, p. 40100a, Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4

[CRW96] Crowston, K.; Williams, M. (1996). Reproduced and emergent genres of communication on the World-Wide Web.

[CRW98] Crowston, K.; Williams, M. (1998). Reproduced and emergent genres of communication on the World-Wide Web.

[DEWE] Dewe, Johan; Karlgren, Jussi; Bretan, Ivan (1998). Assembling a Balanced Corpus from the Internet, In 11th Nordic Conference of Computational Linguistics, Copenhagen, Denmark.

[DEW] Dewdney, Nigel; VanEss-Dykema, Carol; McMillan, Richard (2001). The form is the substance: Classification of genres in text. In ACL Workshop on Human Language Technology and Knowledge Management.

[DIL] Dillon, A.; Gushrowski, B. (2000). Genres and the Web - is the home page the first digital genre? In Journal of the American Society for Information Science, 51(2), 202-205.

[FIN] Finn, A.; Kushmerick, N. (2003). Learning to classify documents according to genre. IJCAI-2003 Workshop on Computational Approaches to Text Style and Synthesis, Acapulco.

[FUR] Furuta, R.; Marshall, C. C. (1996). Genre as Reflection of Technology in the World-Wide Web. Technical report, Hypermedia Research Lab, Department of Computer Science, Texas A&M University.

[GRI] Grieser, Gunter und Fürnkranz, Johannes: Skript zur Vorlesung »Einführung in Maschinelles Lernen und Data Mining«, TU Darmstadt.

[ILL] Illouz, G.; Habert, B.; Folch, H.; Fleury, S.; Heiden, S.; Lafon, P.; Prvost, S. (2000). TyPTex: Generic features for Text Profiler, Content-Based Multimedia Information Access, Paris.

[ISG] de Saint-Georges, Ingrid (1998). Click Here if You Want to Know Who I am. Deixis in Personal Homepages, In Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences-Volume 2 - Volume 2

[JOA2] Joachims, Thorsten (1998). Text categorization with support vector machines: learning with many relevant features. In Proc. 10th European Conference on Machine Learning ECML-98, pages 137-142.

[JOA] Joachims, Thorsten (2001). A Statistical Learning Model of Text Classification for Support Vector Machines. In SIGIR'01, New Orleans.

[KAC] Karlgren, Jussi; Cutting, Douglass (1994). Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. Proc. of COLING94, Kyoto.

[KDD] Böhm, Christian; Ester, Martin; Januzai, Eshref; Kailing, Karin; Kröger, Peer; Sander, Jörg und Shcubert, Matthias (2003). Skript zur Vorlesung »Knowledge Discovery in Databases«. LMU, München.

[KES] Kessler, Brett; Nunberg, Geoffrey; Schütze, Hinrich (1997). Automatic detection of text genre. In Proceedings of the 35th Association of Computational Linguistics (ACL '97), pages 32-38, Madrid, Spain.

[PLU] Plum, G. A.; Cowling, A. (1987). Social constraints on grammatical variables: Tense choice in english. In Steele, Roos und Threadgold, Terry: Language topics. Essays in honour of Michael Halliday, Benjamins, Amsterdam.

[REH] Rehm, Georg (2002). Towards Automatic Web Genre Identification - A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage. In: Proceedings of the Hawaii International Conference on System Sciences. Big Island, Hawaii.

[ROU] Roussinov, D.; Crowston, K.; Nilan, M.; Kwasnik, B. H.; Liu, X.; Cai, J. (2001). Genre-based navigation on the web. In Proceedings of the Thirty-Fourth Hawai'i International Conference on Systems Science (HICSS–34)

[SAN] Santorini, Beatrice (1991). Part-of-Speech Tagging Guidelines for the Penn Treebank Project, Tech. Rep. MS-CIS90 -47, Line Lab 178, University of Pennsylvania, Philadelphia.

[SCH] Schmidt, Helmut: TreeTagger. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

[SCJ] Schneider, Wolf; Raue, Paul-Josef (1998). Handbuch Journalismus, Rowohlt Verlag, Reinbeck bei Hamburg.

[SCU] Schulz, Klaus U. (2003). Nachkorrektur von Ergebnissen einer optischen Charaktererkennun. LMU, München.

[SHE98] Shepherd, Michael; Watters, Carolyn (1998): The Evolution of Cybergenres, In Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences-Volume 2 - Volume 2

[SHE99] Shepherd, Michael; Watters, Carolyn (1999). The Functionality Attribute of Cybergenres, InProceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences-Volume 2 - Volume 2

[STA] Stamatatos, E.; Fakotakis, N.; Kokkinakis, G. (2000). Text Genre Detection Using Common Word Frequencies - Dept. of Electrical and Computer Engineering University of Patras

[STR] Strohmaier, Christian; Ringlstetter, Christoph; Schulz, Klaus U. und Stoyan Mihov (2003). Lexical Postcorrection of OCR-Results: The Web as a Dynamic Secondary Dictionary? In Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 03)

[TOC] Toms, E.G.; Campbell, D.G. (1999). Genre as interface metaphor: Exploiting form and function in digital environments [CD-ROM] (ddgen06.ps). In Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences, Maui. Los Alamitos, CA: IEEE Computer Society.

[TOM] Toms, Elaine G. (2001). Recognizing Digital Genres. In: Bulletin of the American Society for Information Science and Technology, Vol. 27, No. 2

[WAS] Wastholm, P.; Kusma, A.; Megyesi, B. (2005). Using Linguistic Data for Genre Classification. In Proceedings of SAIS-SSLS, Mälardalen University, Västerås

[WEKA] Witten, Ian H. und Frank, Eibe (2005). Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco. (http://www.cs.waikato.ac.nz/~ml/weka/)

[WHI] Whitelaw, Casey; Argamon, Shlomo (2004). Systemic Functional Features in Stylistic Text Classification. At AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design

[WIK] Artikel Vers. In: Wikipedia, Die freie Enzyklopädie. Bearbeitungsstand: 30. Januar 2006, 10:08 UTC. URL: http://de.wikipedia.org/w/index.php?title=Vers&oldid=13178966 (Abgerufen: 14. März 2006, 17:54 UTC)

9 Anhang

9.1 Genres

A. Journalismus

A.1 Kommentar
A.2 Rezension
A.3 Porträt
A.4 Glosse
A.5 Interview und Diskussion
A.6 Nachrichten
A.7 Feature
A.8 Reportage

B. Literatur

B.1 Gedicht
B.2 Prosa
B.3 Drama
B.4 Kurztexte

C. Information/Wissen

C.1 wissenschaftlicher Bericht
C.2 Erklärungen
C.3 Anleitung
C.4 FAQ
C.5 Lexikon
C.6 Zweisprachiges Wörterbuch
C.7 Präsentation, Werbung
C.8 Statisitken
C.9 Code

D. Dokumentation

D.1 Gesetze und Regeln
D.2 Offizieller Bericht
D.3 Protokolle
D.4 Zitate

E. Verzeichnis/Directory

E.1 Personen
E.2 Katalog
E.3 Ressourcen
E.4 Timelines
E.5 Wortlisten
E.6 Sonstige Listen

F. Kommunikation

F.1 Brief/Mail/Rede
F.2 Forum, Gästebuch
F.3 Blog
F.4 Formulare

G. Nichts

H. Kombinationen

9.2 Aufgabenbeschreibung für die Textklassifikation

1. Die Genreliste mit den Definitionen lesen und bei Unklarheiten nachfragen.
2. Vom Internet trennen (da manche Seiten Inhalt nachladen)
3. Die 70 Texte im Korpus-Ordner sortieren. Dabei in die Datei Texte.txt eintragen, welches Genre der Text hat. Gehört ein Text zu mehreren Klassen und ist eine eindeutige Zuordnung nicht möglich, so können diese mit Komma getrennt alle angegeben werden; dabei ist die beste mit einem * zu kennzeichnen.
4. Wenn ein Text ein Genre *enthält*, z.B. wenn in einem Forum Codebeispiele stehen, kann dies zusätzlich angeben und mit einem »X« markiert werden.
5. Falls völlig unklar ist, um was für eine Textart es sich handelt, sollte zumindest dessen Hauptkategorie (Journalismus, Literatur etc.) angegeben werden.
6. Wenn das auch nicht möglich ist, kann der Text übersprungen werden.

Ein paar Beispieleinträge:

file25.html	E.5*, E.1 X-C.1, X-E.6
file1.html	A.1 // evt. Anmerkungen
file3.html	C

9.3 Features

Die folgende Aufzählung stellt alle verwendeten Features dar. Genauere Angaben zu den einzelnen Definitionen finden sich in Kapitel 4.

Struktur & HTML

- Textlänge
- Zeilenanzahl
- Satzanzahl
- durchschnittliche Satzlänge
- Durchschnittliche Zeilenlänge des längsten Gedicht-Blocks
- Länge diese Blocks in Zeilen
- Länge diese Blocks in Zeichen
- Anzahl aufeinanderfolgende Zeilen, die auf eines der Zeichen ; > { } enden, wobei Zeilen mit Kommentaren (/*comment*/, //comment oder #comment) nicht betrachtet werden
- HTML-Zeichenzahl (ohne Inhalte)
- Code-to-Content-Ratio (CCR)
- Überschriften (`<H1>` bis `<H6>`)
- Codetags: `<pre>`, `<xmp>`, `<code>`, `<samp>` und `<font>`-Tags mit Monospace-Schriftart Courier oder CSS-Klasse mit Namen »preformat«, »pre« oder »code«
- Listenelemente (`<li>`)
- Anteil Formularcode an Gesamtlänge
- Länge von Personenlisten (Zeilen und Zeichen)
- Länge von Timelines
- Länge aller bibliographischre Referenzen
- Fehler in der Sortierreihenfolge von...
  - Namenslisten
  - Glossaren
  - Timelines
- Anzahl der sortierten Zeilen in Personenlisten
- Änderung der Sortierung von Timelines

Part of Speech

- Verben
  - Verlaufsform und Gerund (ing-Form)
  - im Präsens
  - im Präsens 3. Person
  - im Simple Past
  - »are« gefolgt von Past Participle
  - »has been«
  - am Satzanfang
- Adjektive
  - positive (→ Wortlisten/adjectives_pos.txt)
  - negative (→ Wortlisten/adjectives_neg.txt)
  - positive + negative

  - neutrale (→ Wortlisten/adjectives_neut.txt)
  - Verhältnis positiver zu negativer Adjektive
  - Verhältnis positiver + negativer zu allen Adjektiven
- Artikel
  - definite
  - indefinite
- Personalpronomen
  - 1. Person singular
  - 3. Person feminin singular
  - 3. Person maskulin singular
  - Verhältnis aus beiden
  - 3. Person neutrum singular
  - 1. Person plural
  - 3. Person plural
  - 1. Person gesamt
  - 2. Person gesamt
  - 3. Person gesamt
  - alle
- Konjunktionen
- Negationen

Vokabular, Wörter und Patterns

- Vorkommen der 200 000 häufigsten engischen Wörtern (→ Wortlisten/general_english.txt)
- Vorkommen anderer Wörter
- Verhältnis von beiden
- Altertümliche Wörter (→ Wortlisten/archaic.txt)
- Wörter und Phrasen zur Zustimmung/Ablehnung
- Konditional (would, should, could, will)
- Synonyme für Sprechen (→ Wortlisten/speak.txt)
- Kausalwörter (→ Wortlisten/kausal.txt)
- Schimpfwörter
- Verallgemeinernde Wörter
- Vage Wörter
- Argumentierende Wörter
- Informelle Wörter und einzelne Wörter in Anführungszeichen
- Verben der Wahrnehmung (→ Wortlisten/percieve.txt)
- Synonyme für Sprechen (→ Wortlisten/speak.txt)
- Zahlen
- Ordinalzahlen
- Datumsangaben
- Datumsangaben inklusive deiktischer Zeitangaben
- Vergangenheits-Keywords (»century« und Jahreszahlen vor 1980)
- Deiktische Zeitangaben
- Deiktische Ortsangaben
- Summe aus beiden
- Maßangaben

- Preisangaben
- Kontraktionen (mit und ohne »n't«)
- Emoticons (→Wortlisten/emoticons.txt)
- Akronyme (→Wortlisten/akronyme.txt)
- Begrüßungen am Textanfang
- Offizielle Begrüßungen am Textanfang
- Verabschiedungen am Textende
- Fehlermeldungen von Skriptsprachen oder Webhostern
- Andere Fehlermeldungen
- Wörter in Anführungszeichen
- Auf -ing oder -ly endende Wörter oder Regieanweisungen in Klammern
- Zeichen-Wiederholungen
- Alphabet (wobei Buchstaben-Bereiche wie A-E auch erkannt werden)
- Typische Variablennamen
- Namen, mit mehr oder weniger strenger Fiterung von englischen Wörtern (→Wortlisten/names.txt)
- Lebewesen (→Wortlisten/animals.txt, humans.txt)
- Anzahl unterschiedlicher Vornamen
- Anzahl unterschiedlicher Nachnamen
- Vorkommen des häufigsten Vornamens
- Vorkommen des häufigsten Nachnamens
- Länder- und Städtenamen (→Wortlisten/cities.txt, countries.txt)

Satzzeichen:

- Fragezeichen
- Ausrufezeichen
- Kommas
- Doppelpunkte
- Anführungszeichen
- Prozentzeichen

HTML-Patterns:

- Text-Formularelemente (`<input type=text>` oder `<textarea>`)
- File-Links
- Dokumenten-interne Links
- Website-interne Links
- Externe Links
- ein Tag enthält `class=xl...` : erstellt mit Excel
- Emoticon-Bilder (Smilies)
- `<link>`-Tag für RSS

Keywords:

- Typische Bigramme für wissenschaftliche Texte mit »we« und »our« (→Wortlisten/science_bigramme.txt)
- Kennzeichnende Patterns für Blog-Anbieter
- Schlüsselwörter in Porgrammiersprachen, die keine echten Wörter sind (→Wortlisten/code.txt)

- Negative Keywords für C8
- Keywords für: A.2, A.3, A.4, A.5, B.2, B.3, C.4, C.5, C.7, D.1, D.3, F1, F.2

Patterns in Struktur/Formatierung

- Zeilen mit Fragezeichen am Ende
- Anzahl solcher Zeilen mit typischen Fragewörtern (Wortlisten/questions.txt)
- Zeilen mit Doppelpunkt nach den ersten 1-30 Zeichen
- solche Zeilen mit Pronomen 1. oder 2. Person singular
- Anzahl der Interviewpartner (unterschiedliche Wörter vor Doppelpunkt)
- Anzahl der Interviewpartner mit mehr als einem Gesprächsbeitrag
- Überschriften für wissenschaftliche Texte (2 Stufen)
- Überschriften für Rezepte und Anleitungen
- Pronomen 3. Person plural in Großbuchstaben
- Namen in Großbuchstaben
- typische Regieanweisungen zu Sprechertexten in Großbuchstaben
- typische Zeit- und Ortsangaben in Regieanweisungen in Großbuchstaben
- nach amerikanischem Standard formatierte Zahlen in Tabellenzellen
- aufeinanderfolgende solche Patterns
- Datumsangaben, die nicht in Links stehen
- C4-Keywords, die nicht in Links stehen
- F3-Keywords in Links
- »posted:« nach schließenden Tags

In Gedicht-Block:

- Wörter in Großbuchstaben
- Sonderzeichen
- Zahlen
- Doppelpunkte am Zeilenende
- typische in Sätzen vorkommende Wörter
- füllt der Block den kompletten Bereich aus, falls er in `<pre>`-Tags steht?

In bibliographischer Referenz:

- Nomen
- Verben
- Wochentage

Im längsten Formular:

- Anzahl der Text-Formularfelder
- Anzahl der Select-Elemente

Kombinationen

- Lebewesen + Pronomen 3. Person + Anführungszeichen
- Verhältnis Namen (incl. Städte, Länder etc.) zu Wörtern in General English
- Pronomen 1. Person + Anführungszeichen (direkte Rede)
- Verhältnis Kontraktionen zu Anführungszeichen
- Vorkommen des häufigsten Vornamens + 2, falls dieser in einer Überschrift am Textanfang vorkommt + Anzahl »he« bzw. »she«

- Vorkommen des häufigsten Nachnamens + 1, falls dieser in einer Überschrift am Textanfang vorkommt + häufister Vorname + Anzahl »he« bzw. »she«
- Ordinalzahlen am Zeilenanfang + Article/Section gefolgt von Nummerierung
- Fehler in der Sortierung von Timelines, abhängig von absoluter und relativer Listenlänge
- gewichtete Summe aus Emoticons, Acronymen, Emoticon-Bildern, Zeichen-Wiederholungen und F2-Keywords
- Böse Sprache: Schimpfwörter + negative Adjektive - positive Adjektive + verallgemeinernde Wörter
- Argumentierende Sprache: 2 · Anzahl der Fragen + Kausal + Konditional + argumentierende Wörter + verallgemeinernde Wörter + Negationen
- Lockere Sprache: Kontraktionen + verallgemeinerd + vage Wörter + informelle Wörter

9.4 Filter-Regeln

R.A1 wenn A.1, dann nicht A.2
R.A2 wenn A.2, dann nicht A.8, C.2, F.1
R.A3 wenn A.3, dann nicht A.1, A.2, A.4
R.A4 wenn A.4, dann nicht A.2, B.2, C.7, F.1
R.A5 wenn A.5, dann nicht A.2, A.4, A.8, B.2, C.2, C.4, C.7, E.1, F.1
R.A6 wenn A.6, dann nicht A.8
R.A7 wenn A.7, dann nicht A.1, A.4, A.6, A.8, C.7, F.1
R.A8 wenn A.8, dann nicht B.2, F.1

R.B1 wenn B.1, dann nichts außer B.2, C.5, F.2, F.3, F.4
R.B2 wenn B.2, dann nicht B.1
R.B3 wenn B.3, dann nichts Anderes

R.C1 wenn C.1, dann nicht A.2, A.7, A.8, B.2, keines der Teilgenres: C.4, C.9, E.1, E.3
R.C2 wenn C.2, dann nicht A.6
R.C3 wenn C.3, dann nicht A.1
R.C4 wenn C.4, dann nicht A.1, B.2, C.2, C.3, C.7, F.1
R.C5 wenn C.5, dann nicht A.1, A.2, A.4, B.1, C.2, C.4, D.1, E.1, F.1
R.C6 wenn C.6, dann nichts außer C.9
R.C9 wenn enthält C.9, dann nicht A, B, C.6, D.1, F.1

R.D1 wenn D.1, dann nicht A.2, C.2
R.D2 wenn D.2, dann nicht A.1, A.7, C.2, C.7
R.D3 wenn D.3, dann nicht D.1, D.2

R.E1 wenn E.1, dann nicht B.2
R.E2 wenn E.2, dann nicht A.6, C.3, C.4, C.7, E1, E3, E4, F.1
R.E3 wenn E.3, dann nicht A.4, C.1, C.3, D.1
R.E3' wenn enthält E.3, dann nicht C.7
R.E4 wenn E.4, dann nicht A.6

R.F2 wenn F.2, dann nicht A.3, A.4, C.3, C.4, E.4, F.3
R.F3 wenn F.3, dann nicht A.2, A.3, A.4, A.5, A.8, B.2, E.3, E.4, F.1
R.F4 wenn F.4, dann nichts außer G

R.G wenn G, dann nichts Anderes

Tabelle 9.1: Filterreglen

9.5 Recall und Precision

Genre	Richtig		Falsch	Recall	Precision	
	alle	Orig.			alle	Orig.
Gesamt	454	381	240	60,5	65,4	54,9
A. Journalismus	92	67	91	41,9	50,3	36,6
A.1 Kommentar	10	6	10	30,0	50,0	30,0
A.2 Rezension	11	10	15	50,0	42,3	38,5
A.3 Porträt	13	11	9	55,0	59,1	50,0
A.4 Glosse	7	2	19	10,0	28,0	7,7
A.5 Interview und Diskussion	14	14	4	70,0	77,8	77,8
A.6 Nachrichten	12	7	13	35,0	48,0	28,0
A.7 Feature	10	7	10	35,0	50,0	35,0
A.8 Reportage	15	10	11	50,0	57,7	38,5
B. Literatur	37	37	19	61,7	66,1	66,1
B.1 Gedicht	14	14	3	70,0	82,4	82,4
B.2 Prosa	15	15	14	75,0	51,7	51,7
B.3 Drama	8	8	2	40,0	80,0	80,0
C. Information/Wissen	132	113	54	66,5	71,0	60,8
C.1 wissenschaftlicher Bericht	15	14	0	70,0	100,0	93,3
C.2 Erklärungen	10	8	15	40,0	40,0	32,0
C.3 Anleitung	20	16	5	80,0	80,0	64,0
C.4 FAQ	15	14	6	70,0	71,4	66,7
C.5 Lexikon	16	15	6	75,0	72,7	68,2
C.6 Zweisprachiges Wörterbuch	8	8	5	80,0	61,5	61,5
C.7 Präsentation, Werbung	13	9	16	45,0	44,8	31,0
C.8 Statisitken	18	12	1	60,0	94,7	63,2
C.9 Code	17	17	0	85,0	100,0	100,0
D. Dokumentation	36	32	9	53,3	80,0	71,1
D.1 Gesetze und Regeln	10	10	2	50,0	83,3	83,3
D.2 Offizieller Bericht	13	9	3	45,0	81,3	56,3
D.3 Protokolle	13	13	4	65,0	76,5	76,5
E. Verzeichnis/Directory	75	57	22	71,3	77,3	58,8
E.1 Personen	14	12	1	69,0	93,3	80,0
E.2 Katalog	19	17	0	85,0	100,0	89,5
E.3 Ressourcen	22	15	6	75,0	81,5	55,6
E.4 Timelines	20	13	15	65,0	57,1	37,1
F. Kommunikation	62	55	45	57,5	57,9	42,1
F.1 Brief/Mail/Rede	7	7	34	35,0	17,1	17,1
F.2 Forum, Gästebuch	23	16	6	80,0	76,7	53,3
F.3 Blog	14	14	1	70,0	93,3	93,3
F.4 Formulare	18	18	4	90,0	81,8	81,8
G. Nichts	20	20	0	100,0	100,0	100,0

Tabelle 9.2: Reacall und Precision für Mehrfachklassifikation

Genre	Richtig		Falsch	Recall	Precision	
	alle	Orig.			alle	Orig.
Gesamt	421	368	141	58,4	74,9	65,5
A. Journalismus	83	65	56	40,6	59,7	46,8
A.1 Kommentar	10	6	5	30,0	66,7	40,0
A.2 Rezension	9	9	8	45,0	52,9	52,9
A.3 Porträt	13	11	6	55,0	68,4	57,9
A.4 Glosse	4	1	11	5,0	26,7	6,7
A.5 Interview und Diskussion	14	14	3	70,0	82,4	82,4
A.6 Nachrichten	11	7	9	35,0	55,0	35,0
A.7 Feature	9	7	6	35,0	60,0	46,7
A.8 Reportage	13	10	8	50,0	61,9	47,6
B. Literatur	36	36	10	60,0	78,3	78,3
B.1 Gedicht	13	13	2	65,0	86,7	86,7
B.2 Prosa	15	15	7	65,0	68,2	68,2
B.3 Drama	8	8	1	40,0	88,9	88,9
C. Information/Wissen	122	109	32	64,1	79,2	70,8
C.1 wissenschaftlicher Bericht	15	14	0	70,0	100,0	93,3
C.2 Erklärungen	8	7	7	35,0	53,3	46,7
C.3 Anleitung	15	14	3	70,0	83,3	77,8
C.4 FAQ	15	14	4	70,0	78,9	73,7
C.5 Lexikon	16	15	5	75,0	76,2	71,4
C.6 Zweisprachiges Wörterbuch	8	8	1	40,0	88,9	88,9
C.7 Präsentation, Werbung	10	8	11	40,0	47,6	38,1
C.8 Statisitken	18	12	1	60,0	94,7	63,2
C.9 Code	17	17	0	85,0	100,0	100,0
D. Dokumentation	35	32	8	53,3	81,4	74,4
D.1 Gesetze und Regeln	10	10	2	50,0	83,3	83,3
D.2 Offizieller Bericht	12	9	2	45,0	85,7	64,3
D.3 Protokolle	13	13	4	65,0	76,5	76,5
E. Verzeichnis/Directory	67	54	12	67,5	84,8	68,4
E.1 Personen	14	12	1	60,0	93,3	80,0
E.2 Katalog	18	17	0	85,0	100,0	94,4
E.3 Ressourcen	17	14	6	70,0	73,9	60,9
E.4 Timelines	18	11	5	55,0	78,3	47,8
F. Kommunikation	58	52	23	65,0	71,6	64,2
F.1 Brief/Mail/Rede	5	5	12	25,0	29,4	29,4
F.2 Forum, Gästebuch	22	16	6	80,0	78,6	57,1
F.3 Blog	13	13	1	65,0	92,9	92,9
F.4 Formulare	18	18	4	90,0	81,8	81,8
G. Nichts	20	20	0	100,0	100,0	100,0

Tabelle 9.3: Reacall und Precision für gefilterte Mehrfachklassifikation

Genre	Richtig		Falsch	Recall	Precision	
	alle	Orig.			alle	Orig.
Gesamt	370	327	101	51,9	78,6	69,4
A. Journalismus	78	63	34	39,4	69,6	56,3
A.1 Kommentar	10	6	1	30,0	90,1	54,5
A.2 Rezension	10	10	4	20,0	71,4	71,4
A.3 Porträt	12	10	5	20,0	70,6	58,8
A.4 Glosse	2	1	3	5,0	40,0	20,0
A.5 Interview und Diskussion	14	14	3	70,0	82,4	82,4
A.6 Nachrichten	9	6	6	30,0	60,0	40,0
A.7 Feature	10	7	7	35,0	58,8	41,2
A.8 Reportage	11	9	5	45,0	68,8	56,3
B. Literatur	32	32	10	53,3	76,2	76,2
B.1 Gedicht	12	12	2	60,0	85,7	85,7
B.2 Prosa	12	12	7	60,0	63,2	63,2
B.3 Drama	8	8	1	40,0	88,9	88,9
C. Information/Wissen	98	88	28	51,8	77,8	69,8
C.1 wissenschaftlicher Bericht	5	4	0	20,0	100,0	80,0
C.2 Erklärungen	6	6	4	30,0	60,0	60,0
C.3 Anleitung	15	14	2	70,0	88,2	82,4
C.4 FAQ	12	12	4	60,0	75,0	75,0
C.5 Lexikon	14	14	6	70,0	70,0	70,0
C.6 Zweisprachiges Wörterbuch	5	5	0	25,0	100,0	100,0
C.7 Präsentation, Werbung	9	7	11	35,0	45,0	35,0
C.8 Statisitken	18	12	1	60,0	94,7	63,2
C.9 Code	14	14	0	70,0	100,0	100,0
D. Dokumentation	33	30	8	50,0	80,5	73,2
D.1 Gesetze und Regeln	10	10	2	50,0	83,3	83,3
D.2 Offizieller Bericht	10	7	2	35,0	83,3	58,3
D.3 Protokolle	13	13	4	65,0	76,5	76,5
E. Verzeichnis/Directory	53	45	7	56,3	88,3	75,0
E.1 Personen	8	7	1	35,0	88,9	77,8
E.2 Katalog	19	17	0	85,0	100,0	89,5
E.3 Ressourcen	18	15	5	75,0	78,3	65,2
E.4 Timelines	8	6	1	30,0	88,9	66,7
F. Kommunikation	56	49	14	61,3	80,0	70,0
F.1 Brief/Mail/Rede	4	4	6	20,0	40,0	40,0
F.2 Forum, Gästebuch	23	16	6	80,0	79,3	55,2
F.3 Blog	13	13	0	65,0	100,0	100,0
F.4 Formulare	16	16	2	80,0	88,9	88,9
G. Nichts	20	20	0	100,0	100,0	100,0

Tabelle 9.4: Reacall und Precision für Auswahl nach F1-Wert

Genre	Richtig		Falsch	Recall	Precision	
	alle	Orig.			alle	Orig.
Gesamt	379	340	92	54,0	80,5	72,2
A. Journalismus	76	61	31	38,1	71,0	57,0
A.1 Kommentar	10	6	2	30,0	83,3	50,0
A.2 Rezension	8	8	3	40,0	72,7	72,7
A.3 Porträt	10	10	3	50,0	76,9	76,9
A.4 Glosse	3	1	4	5,0	42,9	14,3
A.5 Interview und Diskussion	13	13	3	65,0	81,3	81,3
A.6 Nachrichten	9	6	6	30,0	60,0	40,0
A.7 Feature	9	7	4	35,0	69,2	53,8
A.8 Reportage	14	10	6	50,0	70,0	50,0
B. Literatur	32	32	9	53,3	78,0	78,0
B.1 Gedicht	12	12	2	60,0	85,7	85,7
B.2 Prosa	12	12	6	60,0	66,7	66,7
B.3 Drama	8	8	1	40,0	88,9	88,9
C. Information/Wissen	102	94	25	55,3	80,3	74,0
C.1 wissenschaftlicher Bericht	9	8	0	40,0	100,0	88,9
C.2 Erklärungen	8	7	6	35,0	57,1	50,0
C.3 Anleitung	14	13	2	65,0	87,5	81,3
C.4 FAQ	13	13	2	65,0	86,7	86,7
C.5 Lexikon	15	14	5	70,0	75,0,0	70
C.6 Zweisprachiges Wörterbuch	8	8	1	40,0	88,9	88,9
C.7 Präsentation, Werbung	8	6	9	30,0	47,1	35,3
C.8 Statisitken	10	8	0	40,0	100,0	80,0
C.9 Code	17	17	0	85,0	100,0	100,0
D. Dokumentation	34	31	6	51,7	85,0	77,5
D.1 Gesetze und Regeln	10	10	2	50,0	83,3	83,3
D.2 Offizieller Bericht	11	8	2	40,0	84,6	61,5
D.3 Protokolle	13	13	2	65,0	86,7	86,7
E. Verzeichnis/Directory	61	51	6	63,8	91,0	76,1
E.1 Personen	11	10	0	50,0	100,0	90,9
E.2 Katalog	18	17	0	85,0	100,0	94,4
E.3 Ressourcen	15	14	2	70,0	88,2	82,4
E.4 Timelines	17	10	4	50,0	81,0	47,6
F. Kommunikation	54	51	15	63,8	78,3	73,9
F.1 Brief/Mail/Rede	4	4	6	20,0	40,0	40,0
F.2 Forum, Gästebuch	19	16	6	80,0	76,0	64,0
F.3 Blog	13	13	1	65,0	92,9	92,9
F.4 Formulare	18	18	2	90,0	90,0	90,0
G. Nichts	20	20	0	100,0	100,0	100,0

Tabelle 9.5: Reacall und Precision für Auswahl nach Auswertungsreihenfolge

9.6 Accuracy und Classification Error

Genre	Richtig aus Klasse		Falsch	Accuracy		Error
	alle	in Orig.		alle	Orig.	
Gesamt	454	381	240	65,4	54,9	34,6
A. Journalismus	86	67	76	53,1	41,4	46,9
A.1 Kommentar	11	6	5	68,8	37,5	31,2
A.2 Rezension	13	10	6	68,4	52,6	31,6
A.3 Porträt	11	11	7	61,6	61,1	38,9
A.4 Glosse	5	2	8	38,5	15,4	61,5
A.5 Interview und Diskussion	15	14	15	50,0	46,7	50,0
A.6 Nachrichten	9	7	5	64,3	50,0	35,7
A.7 Feature	11	7	18	37,9	24,1	62,1
A.8 Reportage	11	10	12	47,8	43,5	52,2
B. Literatur	40	37	21	65,6	60,7	34,4
B.1 Gedicht	16	14	9	64,0	56,0	36,0
B.2 Prosa	16	15	4	80,0	75,0	20,0
B.3 Drama	8	8	8	50,0	50,0	50,0
C. Information/Wissen	136	113	62	68,7	57,1	31,3
C.1 wissenschaftlicher Bericht	19	14	16	54,3	40,0	45,7
C.2 Erklärungen	9	8	9	50,0	44,4	50,0
C.3 Anleitung	17	16	1	94,4	88,9	5,6
C.4 FAQ	14	14	10	58,3	58,3	41,7
C.5 Lexikon	15	15	10	60,0	60,0	40,0
C.6 Zweisprachiges Wörterbuch	10	8	2	83,3	66,7	16,7
C.7 Präsentation, Werbung	10	9	4	71,4	64,3	28,6
C.8 Statisitken	18	12	4	81,8	54,5	18,2
C.9 Code	24	17	6	80,0	56,7	20,0
D. Dokumentation	38	32	17	69,1	58,2	30,9
D.1 Gesetze und Regeln	10	10	7	58,8	58,8	41,2
D.2 Offizieller Bericht	14	9	8	63,6	40,9	36,4
D.3 Protokolle	14	13	2	87,5	81,3	12,5
E. Verzeichnis/Directory	69	57	17	80,2	66,3	19,8
E.1 Personen	19	12	3	86,4	54,5	13,6
E.2 Katalog	21	17	3	87,5	70,8	12,5
E.3 Ressourcen	15	15	1	93,8	93,8	6,2
E.4 Timelines	14	13	10	58,3	54,2	41,7
F. Kommunikation	65	55	46	58,6	49,5	41,4
F.1 Brief/Mail/Rede	10	7	7	58,8	41,2	41,2
F.2 Forum, Gästebuch	17	16	8	68,0	64,0	32,0
F.3 Blog	18	14	23	43,9	34,1	56,1
F.4 Formulare	20	18	8	71,4	64,3	28,6
G. Nichts	20	20	1	95,2	95,2	4,8

Tabelle 9.6: Accuracy und Classification Error für Merhfachklassifikation

Genre	Richtig aus Klasse		Falsch	Accuracy		Error
	alle	in Orig.		alle	Orig.	
Gesamt	421	368	141	75,2	65,6	24,8
A. Journalismus	82	65	50	62,1	49,2	37,9
A.1 Kommentar	11	6	5	68,8	37,5	31,2
A.2 Rezension	11	9	5	68,8	56,3	31,2
A.3 Porträt	11	11	4	73,3	73,3	26,7
A.4 Glosse	4	1	7	36,4	9,1	63,6
A.5 Interview und Diskussion	15	14	5	75,0	70,0	25,0
A.6 Nachrichten	9	7	4	69,2	53,8	30,8
A.7 Feature	10	7	9	52,6	36,8	47,4
A.8 Reportage	11	10	11	50,0	45,5	50,0
B. Literatur	39	36	9	83,0	76,6	17,0
B.1 Gedicht	15	13	1	93,8	81,3	6,2
B.2 Prosa	16	15	4	80,0	75,0	20,0
B.3 Drama	8	8	4	66,7	66,7	33,3
C. Information/Wissen	126	109	34	79,4	68,1	20,6
C.1 wissenschaftlicher Bericht	18	14	7	72,0	56,0	28,0
C.2 Erklärungen	8	7	9	47,1	41,2	52,9
C.3 Anleitung	15	14	1	93,8	87,5	6,2
C.4 FAQ	14	14	3	82,4	82,4	17,6
C.5 Lexikon	15	15	6	71,4	71,4	28,6
C.6 Zweisprachiges Wörterbuch	8	8	0	100,0	100,0	0,0
C.7 Präsentation, Werbung	8	8	4	66,7	66,7	33,3
C.8 Statisitken	18	12	4	81,8	54,5	18,2
C.9 Code	22	17	0	100,0	77,3	0,0
D. Dokumentation	36	32	16	69,2	61,5	30,8
D.1 Gesetze und Regeln	10	10	6	62,5	62,5	37,5
D.2 Offizieller Bericht	12	9	8	60,0	45,0	40,0
D.3 Protokolle	14	13	2	87,5	81,3	12,5
E. Verzeichnis/Directory	63	54	13	82,9	71,1	17,1
E.1 Personen	19	12	3	86,4	54,5	13,6
E.2 Katalog	18	17	2	90,0	85,0	10,0
E.3 Ressourcen	14	14	1	93,3	93,3	6,7
E.4 Timelines	12	11	7	63,2	57,9	36,8
F. Kommunikation	55	52	19	74,3	70,3	25,7
F.1 Brief/Mail/Rede	8	5	6	57,1	35,7	42,9
F.2 Forum, Gästebuch	16	16	3	84,2	84,2	15,8
F.3 Blog	13	13	9	59,1	59,1	40,9
F.4 Formulare	18	18	1	94,7	94,7	5,3
G. Nichts	20	20	0	100,0	100,0	0,0

Tabelle 9.7: Accuracy und Classification Error für gefilterte Mehrfachklassifikation

Genre	Richtig aus Klasse		Falsch	Accuracy		Error
	alle	in Orig.		alle	Orig.	
Gesamt	370	327	101	78,6	69,4	21,4
A. Journalismus	76	63	34	69,1	57,3	30,9
A.1 Kommentar	10	6	4	71,4	42,9	28,6
A.2 Rezension	12	10	3	80,0	66,7	20,0
A.3 Porträt	10	10	5	66,7	66,7	33,3
A.4 Glosse	3	1	7	30,0	10,0	70,0
A.5 Interview und Diskussion	14	14	3	82,4	82,4	17,6
A.6 Nachrichten	8	6	3	72,7	54,5	27,3
A.7 Feature	9	7	4	69,2	53,8	30,8
A.8 Reportage	10	9	5	66,7	60,0	33,3
B. Literatur	35	32	8	81,4	74,4	18,6
B.1 Gedicht	14	12	1	93,3	80,0	6,7
B.2 Prosa	13	12	3	81,3	75,0	18,7
B.3 Drama	8	8	4	66,7	66,7	33,3
C. Information/Wissen	101	88	29	77,7	67,7	22,3
C.1 wissenschaftlicher Bericht	7	4	9	43,8	25,0	56,2
C.2 Erklärungen	7	6	9	43,8	37,5	56,2
C.3 Anleitung	14	14	1	93,3	93,3	6,7
C.4 FAQ	12	12	3	80,0	80,0	20,0
C.5 Lexikon	14	14	3	82,4	82,4	17,6
C.6 Zweisprachiges Wörterbuch	7	5	1	87,5	62,5	12,5
C.7 Präsentation, Werbung	8	7	2	80,0	70,0	20,0
C.8 Statisitken	14	12	1	93,3	80,0	6,7
C.9 Code	18	14	0	100,0	77,8	0,0
D. Dokumentation	34	30	11	75,6	66,7	24,4
D.1 Gesetze und Regeln	10	10	6	62,5	62,5	37,5
D.2 Offizieller Bericht	11	7	5	68,8	43,8	31,2
D.3 Protokolle	13	13	0	100,0	100,0	0,0
E. Verzeichnis/Directory	51	45	9	85,0	75,0	15,0
E.1 Personen	11	7	2	84,6	53,8	15,4
E.2 Katalog	18	17	0	100,0	94,4	0,0
E.3 Ressourcen	15	15	0	100,0	100,0	0,0
E.4 Timelines	7	6	7	50,0	42,9	50,0
F. Kommunikation	53	49	10	84,1	77,8	15,9
F.1 Brief/Mail/Rede	6	4	5	54,6	36,4	45,4
F.2 Forum, Gästebuch	16	16	0	100,0	100,0	0,0
F.3 Blog	13	13	4	76,5	76,5	23,5
F.4 Formulare	18	16	1	94,7	84,2	5,3
G. Nichts	20	20	0	100,0	100,0	0,0

Tabelle 9.8: Accuracy und Classification Error für Auswahl nach F1-Wert

Genre	Richtig aus Klasse		Falsch	Accuracy		Error
	alle	in Orig.		alle	Orig.	
Gesamt	379	340	92	80,5	72,2	19,5
A. Journalismus	76	61	33	69,7	56,0	30,3
A.1 Kommentar	10	6	4	71,4	42,9	28,6
A.2 Rezension	11	8	4	73,3	53,3	26,7
A.3 Porträt	10	10	4	71,4	71,4	28,6
A.4 Glosse	4	1	6	40,0	10,0	60,0
A.5 Interview und Diskussion	14	13	3	82,4	76,5	17,6
A.6 Nachrichten	8	6	3	72,7	54,5	27,3
A.7 Feature	9	7	4	69,2	53,8	30,8
A.8 Reportage	10	10	5	66,7	66,7	33,3
B. Literatur	35	32	8	81,4	74,4	18,6
B.1 Gedicht	14	12	1	93,3	80,0	6,7
B.2 Prosa	13	12	3	81,3	75,0	18,7
B.3 Drama	8	8	4	66,7	66,7	33,3
C. Information/Wissen	106	94	25	80,9	71,8	19,1
C.1 wissenschaftlicher Bericht	12	8	4	75,0	50,0	25,0
C.2 Erklärungen	8	7	8	50,0	46,7	50,0
C.3 Anleitung	14	13	1	93,3	86,7	6,7
C.4 FAQ	13	13	2	86,7	86,7	13,3
C.5 Lexikon	14	14	3	82,4	82,4	17,6
C.6 Zweisprachiges Wörterbuch	8	8	0	100,0	100,0	0,0
C.7 Präsentation, Werbung	7	6	4	63,6	54,5	36,4
C.8 Statisitken	12	8	3	80,0	53,3	20,0
C.9 Code	18	17	0	100,0	94,4	0,0
D. Dokumentation	34	31	11	75,6	68,9	24,4
D.1 Gesetze und Regeln	10	10	6	62,5	62,5	37,5
D.2 Offizieller Bericht	11	8	5	68,8	50,0	31,2
D.3 Protokolle	13	13	0	100,0	100,0	0,0
E. Verzeichnis/Directory	54	51	6	90,0	85,0	10,0
E.1 Personen	11	10	2	84,6	76,9	15,4
E.2 Katalog	18	17	0	100,0	94,4	0,0
E.3 Ressourcen	14	14	1	93,3	93,3	6,7
E.4 Timelines	11	10	3	78,6	71,4	21,4
F. Kommunikation	54	51	9	85,7	81,0	14,3
F.1 Brief/Mail/Rede	7	4	4	63,6	36,4	36,4
F.2 Forum, Gästebuch	16	16	0	100,0	100,0	0,0
F.3 Blog	13	13	4	76,5	76,5	23,5
F.4 Formulare	18	18	1	94,7	94,7	5,3
G. Nichts	20	20	0	100,0	100,0	0,0

Tabelle 9.9: Accuracy und Classification Error für Auswahl nach Auswertungsreihenfolge

9.7 Konfusionsmatrix für Mehrfachklassifikation

	A.1	A.2	A.3	A.4	A.5	A.6	A.7	A.8	B.1	B.2	B.3	C.1	C.2	C.3	C.4	C.5	C.6	C.7	C.8	C.9	D.1	D.2	D.3	E.1	E.2	E.3	E.4	F.1	F.2	F.3	F.4	G
A.1	6																1											2				
A.2	1	9		2	1	1	3	1				2	1	1	1	1					1							1				
A.3			11	1	1	1	3	1	1	3				1	1								1			1			1			
A.4	1	2		2	4		2	6	1			1			1													1	3			
A.5				2	14										2																	
A.6	3			1		7										1		1								2						1
A.7			1			2	7			1		4				1					1	1	1									
A.8	3	1	1	2			3	10		1			1									1						1				
B.1									14		1		1			1																
B.2								1		15	1		2		2														4			
B.3				1					1	1	8																					
C.1							1					14										4										
C.2		1	1		1		3					1	8			2		2			1											
C.3												1	2	14				1		3						1		1				
C.4	1				1			1				1			14			1	1	1												
C.5			1									4				15	1					1										
C.6																1	8			4							1					
C.7		1					4	2					1		1			9						2				1				
C.8															1				12								1					
C.9															1					17									4			
D.1		1			1								1					1			10											
D.2						1						5										9										
D.3												2				1						1	13									
E.1																		2						12						1		
E.2	1																								17					2		
E.3	1					2								1									1		2	15						
E.4					1				1				1	1				1				1					13					
F.1		3		4	5	1			5					2	1													7				
F.2									2		1								4										16			
F.3																1														14		
F.4													1												1						18	
G																																20

Abbildung 9.1: Konfusionsmatrix. Spalten: *ist* Klasse; Reihen: *erkannt als* Klasse

9.8 Vergleich der automatischen Klassifikatoren

Genre	Naive Bayes	J48-Tree	k-NN	SVM
A. Journalismus				
A.1 Kommentar	24,6	23,8	16,2	22,7
A.2 Rezension	31,8	24,4	32,7	45,9
A.3 Porträt	26,7	38,7	38,9	58,8
A.4 Glosse	20,8	23,3	11,4	14,3
A.5 Interview und Diskussion	29,4	32,0	40,9	50,0
A.6 Nachrichten	52,4	15,0	22,6	36,7
A.7 Feature	34,1	6,1	19,4	25,0
A.8 Reportage	48,1	30,2	12,8	32,6
B. Literatur				
B.1 Gedicht	60,4	52,4	25,8	66,7
B.2 Prosa	76,5	31,8	46,5	69,8
B.3 Drama	82,9	61,9	78,9	90,0
C. Information/Wissen				
C.1 wissenschaftlicher Bericht	68,4	48,3	32,0	63,2
C.2 Erklärungen	0,0	0,0	17,0	13,0
C.3 Anleitung	57,1	58,5	40,0	51,2
C.4 FAQ	63,2	47,6	14,8	70,0
C.5 Lexikon	29,6	34,8	45,8	52,9
C.6 Zweisprachiges Wörterbuch	15,4	10,0	0,0	33,3
C.7 Präsentation, Werbung	11,8	17,1	16,7	50,0
C.8 Statisitken	43,8	19,0	16,2	37,8
C.9 Code	42,9	44,4	59,5	70,3
D. Dokumentation				
D.1 Gesetze und Regeln	44,4	35,7	28,6	46,7
D.2 Offizieller Bericht	22,2	22,2	10,9	10,5
D.3 Protokolle	76,5	72,0	68,3	85,0
E. Verzeichnis/Directory				
E.1 Personen	40,0	50,0	18,7	50,0
E.2 Katalog	63,0	74,7	33,3	71,8
E.3 Ressourcen	42,4	8,7	35,3	44,4
E.4 Timelines	47,1	40,0	21,1	38,3
F. Kommunikation				
F.1 Brief/Mail/Rede	12,1	37,5	25,0	28,6
F.2 Forum, Gästebuch	50,0	75,7	40,0	62,9
F.3 Blog	60,6	51,6	21,1	43,2
F.4 Formulare	37,7	35,6	54,5	51,6
G. Nichts	85,7	68,2	78,0	71,4

Tabelle 9.10: Vergleich der F1-Werte der automatischen Klassifikatoren

Eidesstattliche Erklärung

Hiermit versichere ich eidesstattlich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Diese Erklärung erstreckt sich auch auf die graphischen Darstellungen. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, habe ich in jedem Fall unter Angabe der Quelle der Entlehnung kenntlich gemacht. Ich versichere, dass die Arbeit noch nicht veröffentlicht oder in einem anderen Prüfungsverfahren als Prüfungsleistung vorgelegt worden ist.

Lebenslauf

06.08.1979	Geboren in Rothenburg ob der Tauber
1985 bis 1989	Besuch der Grundschule
1989 bis 1998	Besuch des Reichsstadt-Gymnasiums Rothenburg, Abitur.
1998 bis 2002	Studium Medien und Informationswesen an der FH Offenburg, Abschluss als Diplom Ingenieur (FH)
	Praktika beim Saarländischen Rundfunk, Schaeffler Interactive, Rickhoff Internet Solutions und im Telekommunikationslabor der Fachhochschule
seit 2002	Studium Computerlinguistik an der LMU München mit den Nebenfächern Informatik und Logik und Wissenschaftstheorie
	Arbeit als studentische Hilfskraft in den Fachbereichen Medieninformatik und Computerlinguistik
Oktober 2003 bis Juni 2004	Angestellte im technischen Dienst beim Medienzentrum der TU München