

Hausarbeit und Bericht: Automatische Autorenerkennung mit Kollokationen als Klassifikationsmerkmale

Armin Schmidt

08/2008

Zusammenfassung

Herkömmliche Ansätze zur automatischen Autorenerkennung, der automatischen Zuordnung eines anonymen Textes zu einem einzelnen Autoren aus einer Gruppe von Kandidaten, beruhen auf stilistischen Merkmalen wie Satzlänge, Häufigkeit von Funktionswörtern, Wortschatz eines Autoren oder auch auf syntaktischen Hinweisen wie Häufigkeit von Nebensätzen, der Länge einzelner Phrasen, etc. Die vorliegende Arbeit baut auf der Hypothese auf, dass Kollokationen, Wortgruppen, deren Bedeutung über die einfache Zusammensetzung der Bedeutungen ihrer Bestandteile hinausgeht, ein wichtiges Merkmal des individuellen Stils eines Autors seien. Zum Prüfen dieser These wurde ein Experiment durchgeführt, dessen Ergebnisse hier vorgestellt werden sollen. Wie sich zeigen wird, muss die Ausgangshypothese abgelehnt werden; obgleich Kollokationen ein Stilmerkmal sein mögen, lassen sie sich mit der hier präsentierten Methode zur automatischen Autorenerkennung nicht erfolgreich einsetzen.

1 Einleitung

Der Begriff *automatische Autorenerkennung* beschreibt die Problemstellung, einem Text, dessen Verfasser unbekannt ist, einen einzelnen Autoren aus einer Gruppe von Kandidaten zuzuordnen. Dies ist eine klassische Klassifikationsaufgabe, deren Methoden und Ergebnisse nicht auf die Domäne der Autorenerkennung beschränkt sein müssen, sondern potentiell auf andere Felder der Textklassifikation angewandt werden können. Ich möchte in diesem Abschnitt die Besonderheiten der Problematik kurz skizzieren und einige bisher verfolgte Ansätze knapp überschlagen.

Das Zuordnen eines anonymen Textes zu einem von mehreren möglichen Verfassern ist beispielsweise in der forensischen Linguistik bei der Analyse von Bekennerschriften von konkreter Bedeutung, aber auch im Bereich der Geschichts- oder Literaturwissenschaft, wo, um eine korrekte Interpretation führen zu können, der Autor eines Dokuments bekannt sein muss. Normalerweise ist die Anzahl der Kandidaten beschränkt auf einige wenige, in der Praxis auf zwei bis fünf [Stamatatos et al., 2001, Abschnitt 1]. Ausschlaggebend für die Performanz eines automatischen Klassifikationsalgorithmus ist oft die Länge eines Dokuments [z.B. Sichel, 1986]; hier unterscheiden sich die Anwendungsdomänen stark voneinander. Kriminelle Schreiben wie sie in der forensischen Linguistik untersucht werden, sind oft sehr kurz und bieten daher relativ wenig Analysematerial, auf das eine Klassifikationsmethode aufbauen kann, welche, und das ist in den hier besprochenen Ansätzen der Fall, einen wohlgeformten Text ohne orthographische Fehler erwartet und aus diesem *stilistische* Merkmale zu extrahieren versucht [Olsson, 2008]. Literarische oder historische Texte hingegen zeigen gerade entgegengesetzte Eigenschaften: sie sind vergleichsweise lang und eher wohlgeformt. In Bezug auf literarische Texte wurde allerdings auf die Problematik der Heterogenität hingewiesen [Stamatatos et al., 2001] - sie bestehen oft auf mehreren Textarten, wie z.B. Dialogen und erzählenden Abschnitten und zielen mitunter gerade auf stilistische Diversität

ab, was für Klassifikationsalgorithmen problematisch sein kann. Stamatatos et al. [2001] weist darauf hin, dass es wünschenswert wäre, entsprechende Abschnitte zuerst zu erkennen und dann diejenigen für die Klassifikation zu nutzen, welche am ehesten den Stil des Autors wiedergeben. Ich bin mir jedoch keiner Arbeit bewusst, die sich dieser Problematik annimmt.

Die hier beschriebenen Experimente streben keine Bearbeitung eines bestimmten Anwendungsszenarios an, sondern verstehen sich als Versuch, eine allgemeine Klassifikationsaufgabe mittels Betrachtung von Kollokationen zu lösen. Dabei steht die Zielsetzung im Vordergrund, ausreichend lange Texte einer bestimmten Domäne (hier: Zeitungsartikel) so zu klassifizieren, dass auch bei einer größeren Anzahl von Kandidaten, in diesem Fall zehn, der richtige ausgewählt wird. Die Arbeitshypothese soll sein, dass Kollokationen als stilistische Merkmale geeignet sind, die Aufgabe befriedigend zu lösen.

Die Hausarbeit ist wie folgt aufgebaut: In Abschnitt 1.1 werde ich knapp einige Ansätze zur automatischen Autorenerkennung beschreiben. Abschnitt 2 begründet die Verwendung von Kollokationen als Klassifikationsmerkmale und führt kurz in den theoretischen Hintergrund ein. In Sektion 3 findet die Beschreibung der Experimente statt. Zunächst werde ich dort die Wahl des ausgewählten Korpus legitimieren. Anschließend stelle ich die Baseline-Methode und deren Ergebnisse vor. Nachfolgend, in Abschnitt 3.3.2, meinen eigenen Ansatz. Eine Fehleranalyse folgt in Abschnitt 3.3.4. Im Anschluss daran werde ich einige Folgeversuche schildern und konzis auswerten. Abschnitt 4 stellt das Fazit, in welchem die Ergebnisse in einem globaleren Maße bewertet, und Hinweise auf zukünftige Aufgaben gegeben werden sollen.

1.1 Überblick über andere Ansätze

Bis vor kurzem nahmen Ansätze zur automatischen Autorenerkennung fast ausschließlich einfache Oberflächenmaße in Gebrauch, zumeist mit Wortformen als kleinste Einheit. Obwohl es in früheren Tagen Versuche gegeben hat, Merkmale auch unterhalb der Wortebene, z.B. in Hinsicht auf die Anzahl Buchstaben [Brinegar, 1963] oder die Anzahl Silben [Fucks, 1952] pro Wort anzuwenden, haben die erfolgreichsten Methoden lexikalische Maße herangezogen. Holmes [1994] beschreibt zwei Trends innerhalb der lexikalischen Methoden: zum einen Maße, die den Wortschatz eines Autors respektive des Textes widerspiegeln. Und zum anderen Maße, welche die Häufigkeit bestimmter Funktionswörter messen. Ersteres lässt sich beispielsweise durch das Zählen von Wörtern, die nur einmal (*Hapax Legomena*) bzw. zweimal (*Dislegomena*) im Text vorkommen, approximieren, was jedoch zu Ergebnissen führt, die stark von der Textlänge abhängen. Häufigkeiten von Funktionswörtern hingegen haben den Nachteil, dass Mengen ausgewählter Funktionswörter schlecht auf verschiedene Gruppen von Autoren angewandt werden können. Die Baseline-Methode meiner Experimente orientiert sich frei an Burrows [1987] und bedient sich der häufigsten Wörter einer Textsammlung, ohne Unterscheidung zwischen Inhalts- und Funktionswörtern.

Ein vielversprechender Ansatz ist Stamatatos et al. [2001], der über die Ebene lexikalischer Maße hinausgeht und Klassifikationsmerkmale auf Basis der Ausgabe eines Satz- und Chunk-Grenzen-Erkenners (SCBD) bestimmt. Dabei unterscheidet er *low-level measures*, wie Satzlänge und Anzahl Interpunktionszeichen, *phrase-level measures*, wie der Länge von Verbal- oder Nominalphrasen, und *analysis-level measures*, welche die Ausgabe des SCBD-Tools näher betrachten und dessen Ungenauigkeiten verarbeiten. Letzteres beinhaltet z.B. die Anzahl von Wörtern, die SCBD nicht erkannt hat und welche somit als potentielle Fremdwörter gehandhabt werden können. Jener Artikel hat auch den Ausschlag für meinen Ansatz gegeben, indem er sagt: „In general, real natural language processing (NLP) (i.e. computational syntactic, semantic, or pragmatic analysis of text) is avoided since current NLP tools do not manage to provide very high accuracy dealing with unrestricted text.“ Die vorliegende Arbeit stellt einen Versuch dar, genau diesen bisher vernachlässigten Bereich anzugehen, indem sie ein semantisches Phänomen hinsichtlich seines Nutzens zur Autorenerkennung untersucht.

2 Kollokationen

Jeder Mensch besitzt eine Menge von Begrifflichkeiten, die er oder sie sich im Laufe des Heranwachsens und später im Studien- und Berufslebens angeeignet hat. Eine besondere Rolle fällt dabei den Idiomen zu, die sich regional zum Teil deutlich unterscheiden und selbst nur innerhalb der eigenen Familie mehr oder weniger häufig gebraucht werden. Diese Eingebung führt mich zu der Annahme, dass Idiome ein potentiell bedeutsames Merkmal für die Autorenschaft eines Textes sein könnte. Idiome in der automatischen Verarbeitung zu verwerten bringt jedoch drei Schwierigkeiten mit sich: Zum einen sind Listen mit idiomatischen Wendungen schwer zu beschaffen. Zweitens sind solche Listen vermutlich nur schlecht auf verschiedene Domänen oder gar einzelne Autoren übertragbar. Drittens kommen reine Idiome leider viel zu selten in den zu klassifizierenden Texten vor, gerade in einer Domäne wie Zeitungsartikel, die ich für meine Experimente genutzt habe. Als Alternative können daher Kollokationen ins Auge gefasst werden. Diese können mit kleinem Aufwand mittels statistischer Methoden aus einer Textsammlung extrahiert werden und schließen zudem, so sie denn vorkommen, potentielle Idiome ein. Kollokationen und Idiome sind zwar vom Standpunkt der lexikalischen Semantik durchaus zu unterscheiden, dieser Unterschied wird jedoch durch die statistische Berechnung weitestgehend ignoriert.

Der Begriff *Kollokation* ist nicht einheitlich definiert. Eingeführt von Firth [1957], bezeichnet er zunächst ein regelhaft erwartbares gemeinsames Auftreten zweier oder mehrerer Wörter. Zentral ist dabei der Gedanke, dass Kollokationen einer semantischen, nicht grammatischen Begründung folgen. Cruse [1986] verfeinert den Begriff, indem er ihn dem des *Idioms* gegenüberstellt und dies im Kontext semantischer Transparenz begründet. Demnach stellt ein Idiom wie *den Löffel abgeben* eine einzige minimale semantische Konstituente dar, in diesem Falle also [sterben]. Eine weitere Einschränkung für Idiome ist, dass sie nicht modifizierbar sind; *?einen Löffel abgeben* ist genauso wenig eine Kollokation mit verwandter Bedeutung wie *?den Löffel und die Gabel abgeben*. Kollokationen hingegen sind semantisch transparent, d.h. sie bestehen aus mehreren semantischen Konstituenten, deren Teile jedoch nur zusammen und im jeweiligen Kontext einen bestimmten semantischen Inhalt formen. Am Beispiel *starker Raucher* lässt sich dies verdeutlichen: der Ausdruck ist semantisch transparent, denn er besteht aus zwei Wörtern, die jeweils eine individuell erschließbare Bedeutung besitzen und zwei semantische Konstituenten repräsentieren, [stark] und [Raucher]. Nichtsdestotrotz selektieren sich ihre Bedeutungen gegenseitig; *stark* im Kontext von *eine starke Frau* drückt einen anderen semantischen Inhalt aus, als in Kontexten wie *starker Raucher*, *starker Trinker*, etc. Kollokationen sind überdies modifizierbar, wie das Beispiel *ein stärkerer Raucher* belegt. Cruse führt weitere Unterschiede in der Klasse der Kollokationen an. So ist z.B. eine gebundene Kollokation (engl. *bound collocation*) eine, bei denen zumindest einer der Teile unikal selektiv ist, d.h. sie kommen in ihrer Bedeutung allein mit einem bestimmten anderen Wort bzw. in einer bestimmten Kollokation vor. Eine gebundene Kollokation wäre beispielsweise das englische *to foot a bill* (dt. *eine Rechnung bezahlen*).

In der Computerlinguistik wird der Begriff der Kollokation normalerweise weniger eingeschränkt definiert. So orientieren sich Manning and Schütze [2003] locker an der Definition von Firth [1957] und beschränken sich weitestgehend auf die Voraussetzung, dass eine Kollokation *beschränkt kompositionell* sei, d.h. die Bedeutung des Ganzen nicht *komplett* aus der Bedeutung der Teile herleitbar sei. Zumindest aber den feineren Unterscheidungsebenen aus Cruse [1986] wird die Behandlung von Kollokationen in Manning and Schütze [2003] nicht gerecht. So unterscheiden die dort vorgestellten Techniken nicht zwischen einfachen und gebundenen Kollokationen, Idiomen oder Eigennamen, extrahieren ggf. auch terminologische Ausdrücke und beschränken sich weitestgehend auf n-Gramme fester Größe, deren Teile also unmittelbar adjazent sein müssen, um als Kollokation aus einer Textsammlung isoliert zu werden. Der Hinweis auf diese Problematik soll keine generelle Kritik an Manning and Schütze [2003] sein; ich denke, dass der dortige Rahmen als auch mögliche Anwendungen einige theoretische Ungenauigkeiten rechtfertigen. Nichtsdestoweniger möchte ich auf sie hinweisen, denn der in der vorliegenden Arbeit eingesetzte Kollokationsbegriff ist linguistisch ebenfalls weitestgehend nachlässig definiert und schließt eben genannte Ungenauigkeiten mit ein.

3 Experimente

3.1 Korpus

Zum Überprüfen der These, Kollokationen seien ein gutes Stilmerkmal anhand dessen ein anonymes Text einem Autoren sicher zugeordnet werden kann, habe ich Artikel aus der Online-Ausgabe der deutschen Zeitung *DIE ZEIT*¹ der Jahre 1999 und 2000 verwendet. *DIE ZEIT* erscheint wöchentlich und richtet sich insbesondere an Leser der Bildungsschicht. Ich halte die Zeitung aus mehreren Gründen gut für die Aufgabe geeignet. Zum einen sind, im Vergleich zu Tageszeitungen, die Artikel der *ZEIT* im Durchschnitt länger. Im Allgemeinen gilt: je länger die Texte, desto erfolgreicher ist die automatische Autorenerkennung [Stamatatos et al., 2001], einfach deshalb, weil aus langen Texten eine größere Anzahl Stilmerkmale extrahiert werden kann. Zum zweiten beschäftigt *DIE ZEIT* nicht ausschließlich professionelle Journalisten, sondern zu einem gewissen Teil Fachleute einzelner Sparten, und veröffentlicht regelmäßig Essays und Meinungsbilder einflussreicher Personen der Gegenwart, wie Schriftsteller, Politiker und Philosophen². Insbesondere Letztere unterliegen weniger stark den von der Zeitung vorgesehenen Stilvorgaben, so dass sich die individuellen Stile der Autoren in den entsprechenden Artikeln realisieren können.

3.1.1 Vorverarbeitung

Da die Artikel des Korpus im HTML-Format vorlagen, waren verschiedene Vorbereitungsschritte notwendig. Zunächst habe ich die Information über die Autoren aller Artikel extrahiert und diejenigen zur Weiterverarbeitung gewählt, die im Laufe der Jahre 1999 bis 2000 mindestens 15 und höchstens 30 Artikel verfasst hatten³. Es ließen sich auch Autoren finden, die insgesamt wesentlich mehr Texte geschrieben hatten, welche jedoch normalerweise zu kurz (< 300 Wörter pro Text) und daher für die Aufgabe der Autorenerkennung ungeeignet waren (Meistens handelte es sich dabei um kurze Nachrichtentexte oder Glossen). Von den verbleibenden Autoren wurden diejenigen 10 für das Experiment gewählt, deren durchschnittliche Textlänge am größten war. Von diesen 10 wiederum wurden die jeweils längsten 18 Texte ausgewählt, denn je länger ein Text, desto mehr Information enthält er, die für die Autorenerkennung wichtig ist. Zur leichteren Weiterverarbeitung habe ich alle Wörter in Kleinschreibung konvertiert und durch Zeilenumbrüche voneinander getrennt, so dass in jeder Zeile genau ein Wort steht. Interpunktionszeichen wurden, so weit es ging, gelöscht⁴. Anschließend wurden die Daten nach dem Zufallsprinzip⁵ in Trainings- und Testmenge mit zwölf beziehungsweise sechs Texten pro Autor geteilt. Algorithmus 1 stellt die Vorverarbeitungsschritte schematisch dar. Tabelle 1 gibt eine Übersicht über das für das Experiment verwendete Datenmaterial. Trainings- und Testmenge zusammen haben eine Größe von 255.424 Wörtern (*tokens*).

¹<http://www.zeit.de>

²Zur Übersicht siehe: <http://www.zeit.de/autoren/A/index.xml>

³Es soll erwähnt sein, dass die Meta-Angaben in den HTML-Dateien, wie z.B. Autor oder Rubrik, keiner einheitlichen Formatierung folgt. Die hier angewandte Extraktionsmethode verwendet reguläre Ausdrücke, so dass der Autor einiger Texte evtl. nicht korrekt identifiziert werden konnte. Für das Experiment wurden nur solche Artikel benutzt, die einem Autoren eindeutig zuzuweisen waren.

⁴Eine Ausnahme hiervon ist der Bindestrich („-“), der auch zusammengesetzte Wörter, insbesondere Nomen, verbindet und erwünscht ist.

⁵Zur zufälligen Auswahl von Dateien habe ich das UNIX-Tool *shuf* benutzt.

Algorithm 1 Vorverarbeitungsschritte

1. Auswahl der Autoren mit n Texten, so dass $15 \leq n \leq 30$
 2. Davon diejenigen 10 Autoren auswählen, deren Texte durchschnittlich am längsten sind
 3. Von jedem der verbleibenden Autor die 18 längsten Texte nehmen
 4. Zu Kleinschreibung konvertieren, Interpunktionszeichen löschen
 5. In Trainings- und Testmengen trennen mit jeweils 12 bzw. 6 Texten pro Autor
-

Tabelle 1: Korpus

Autor	Wörter insgesamt	durchschn. Textlänge	Wortschatz	Themen
a1	21780	1210	7019	Philosophie
a2	18966	1053	5971	Innenpolitik
a3	26783	1487	6532	Politik/Portraits
a4	31140	1730	8357	Wirtschaft
a5	21534	1196	5501	Telekommunikation
a6	29253	1625	7952	Architektur
a7	20885	1160	5705	Technik
a8	30283	1682	7879	Schule/Beruf
a9	29862	1659	8703	Außenpolitik/Rusland
a10	24938	1385	7836	Gesellschaft & Technik
gesamt	255.424	1419	–	

3.2 Baseline

3.2.1 Klassifikation

Die Baseline-Methode klassifiziert einen Text anhand der in ihm vorkommenden Oberflächenformen. Hierzu wurden aus allen 180 Texten die 100 häufigsten Wörter extrahiert. Ihre relative Häufigkeit in den Texten der Trainingsmenge wurden als Merkmale für einen maschinellen Lernalgorithmus verwendet und von diesem einzeln und für die jeweilige Klasse, d.h. den jeweiligen Autor, gewichtet. Anschließend wurden die Texte der Testmenge mittels Weighted Overlap (gewichtete Überlappung) klassifiziert (s.u.).

Zur Traversalion des Korpus sowie zur Berechnung der relativen Worthäufigkeiten habe ich die freie und für die maschinelle Sprachverarbeitung konzipierte Python-Programmbibliothek NLTK [Bird, 2006] eingesetzt. Zur Durchführung der Gewichtung sowie zur eigentlichen Klassifikation habe ich die ebenfalls freie maschinelle Lernsoftware TiMBL [Daelemans et al., 2007] angewandt. TiMBL implementiert mehrere Algorithmen und Metriken basierend auf einem *memory-based* (auch: *example-based*) Lernansatz [Daelemans and van den Bosch, 2005]. Memory-based Learning versucht, neue Situationen mit einer Menge alter gespeicherter Situationen zu vergleichen und wendet keine abstrakten Klassifikationsregeln an. Die Merkmalswerte zum Trainieren und Testen müssen so aufbereitet werden, dass jede Instanz einer Klasse einer Zeile der Eingabedatei entspricht, die wiederum die Merkmalswerte in feststehender Reihenfolge enthält. Jede Zeile stellt also einen Merkmalsvektor dar. Das letzte Feld einer Zeile enthält immer den Klassennamen, in meinem Fall also $a_1 \dots a_{10}$. Bei der Klassifikation wird dann jeder Merkmalsvektor der gespeicherten Beispiele mit dem jeweiligen neuen Eingabevektor verglichen und für letzteren die ähnlichste Klasse anhand eines Ähnlichkeitsmaßes ermittelt.

Für die Experimente wurden durchweg dieselben Einstellungen genutzt, so dass Unterschiede in der Performanz allein auf die Auswahl der Merkmale und nicht auf den gerade in Gebrauch genommenen Lernalgorithmus zurückzuführen sind. Desweiteren habe ich die Standardeinstellungen verwendet, denn die Experimente zielen letztlich lediglich darauf ab, einen Vergleich mehrerer Methoden zu gestatten und sollen keine maximal performanten Ergebnisse erreichen. Eingesetzt wurde dementsprechend die Weighted-Overlap-Metrik für numerische Merkmale.⁶

3.2.2 Ergebnisse

Die Baseline-Methode klassifizierte 38 von 60 Texten der Testmenge richtig, was einer Akkuratheit von rund 0,633 bzw. einem durchschnittlichen Fehler von rund 0,366 entspricht. Tabelle 2 zeigt die entsprechende Konfusionsmatrix. Die Zeilen entsprechen den echten Klassenzugehörigkeiten, die Spalten den von TiMBL vorhergesagten Autoren. In der Spalte ganz rechts sind die gerundeten Klassifikationsfehler pro Autor angegeben.

3.2.3 Fehleranalyse

Es fällt auf, dass die fehlerhaft klassifizierten Texte nicht zufällig gestreut zu sein scheinen. So wurden beispielsweise 7 Texte fälschlicherweise Autor 10 zugeordnet und 6 Texte irrtümlich mit Autor 6 assoziiert (siehe vorletzte Zeile). Den Autoren 1, 3, 6 und 7 jedoch wurden keine Texte fehlerhaft zugewiesen. Betrachtet man noch einmal Tabelle 1, so lässt sich dieser Umstand nicht durch Größe des Vokabular oder durchschnittliche Textlänge einzelner Autoren erklären. Auch eine nähere Betrachtung der Vorkommen einzelner Wörter in den Texten bestimmter Autoren legen hier keine Systematik nahe. Ich vermute daher, dass, entgegen des ersten Eindrucks, die Streuung fehlerhaft zugeordneter Texte zufällig ist (siehe vergleichsweise Abschnitt 3.3.4).

⁶Eine Erläuterung der maschinellen Lernalgorithmen würde den Rahmen dieses Berichts leider sprengen. Für eine übersichtliche Einführung verweise ich auf Daelemans et al. [2007], insbesondere Kapitel 5.

Tabelle 2: Konfusionsmatrix der Baseline-Methode

echte Klasse →	a10	a9	a8	a7	a6	a5	a4	a3	a2	a1	Fehler
a10	4	0	0	0	0	0	1	0	1	0	0,33
a9	1	4	0	0	0	1	0	0	0	0	0,33
a8	1	1	0	0	0	0	2	0	2	0	1
a7	0	0	0	4	0	1	1	0	0	0	0,33
a6	0	0	1	0	4	0	1	0	0	0	0,33
a5	0	0	0	0	0	6	0	0	0	0	0
a4	1	0	1	0	0	0	4	0	0	0	0,33
a3	1	0	0	0	0	0	0	5	0	0	0,16
a2	1	0	0	0	0	1	0	0	4	0	0,33
a1	2	0	0	0	0	0	1	0	0	3	0,5
falsch zugeordnete Texte	7	1	2	0	0	3	6	0	3	0	
	durchschnittl. Fehler:										0,366

3.3 Methode

3.3.1 Berechnung von Kollokationen: der t-Test

Eine häufig angewandte Methode zum Finden von Kollokationen in einem Korpus ist der t-Test. Diese Methode des statistischen Hypothesentests betrachtet den Erwartungswert und die Varianz einer Stichprobe, stellt sie einer Nullhypothese gegenüber und macht eine Aussage darüber, wie sehr die erwartete und die tatsächliche Verteilung voneinander abweichen. Ist die Abweichung groß, so kann die Nullhypothese abgelehnt werden; die Stichprobe muss dann bei einem Signifikanzniveau α aus einer anderen Verteilung stammen. Bei $\alpha = 0.005$ kann die Nullhypothese mit einer Sicherheit von 99,5% abgelehnt werden, wenn $t \geq 2.576$. Dieser Konfidenzwert kann in jedem Statistikbuch nachgeschlagen werden, siehe z.B. Manning and Schütze [2003, S. 309]. Der t-Test stellt eine Abbildung:

$$t = \frac{\bar{x}' - \mu}{\sqrt{\frac{s^2}{N}}}$$

mit dem Erwartungswert der Stichprobe \bar{x}' , dem Erwartungswert der Verteilung μ , der Standardabweichung s^2 , und der Stichprobengröße N dar.

Will man ihn auf die Berechnung von Kollokationen anwenden, so lässt sich der t-Test folgendermaßen erweitern: Ein Textkorpus wird als Sequenz von Bigrammen betrachtet, deren Stichproben Zufallsvariablen darstellen, die den Wert 1 annehmen, wenn das jeweilige Bigramm auftritt, und 0 sonst [Manning and Schütze, 2003]. Unter Verwendung der Maximum-Likelihood-Methode können wir die zur Berechnung des t-Wertes nötigen Parameter schätzen, hier z.B. für das Wortpaar *berliner republik*:

Die Nullhypothese sei, dass die Wörter *berliner* und *republik* unabhängig sind:

$$H_0 : P(\text{berliner}, \text{republik}) = P(\text{berliner}) \times P(\text{republik}) = \frac{40}{255434} \times \frac{69}{255434} \approx 4,23^{-8}$$

Abbildung 1: Die 50 Bigramme mit den höchsten t-Werten:

in der	mit den
in den	zum beispiel
für die	in einem
mit dem	an die
auf dem	mehr als
nicht mehr	dass die
vor allem	bei der
aus dem	ist es
nicht nur	am ende
für den	ein paar
- und	mit der
von der	gar nicht
an der	ist das
über die	wenn sie
gibt es	gegen die
in deutschland	sie sich
auf den	an den
das ist	in einer
auf die	die meisten
aus der	nur noch
mit einem	von den
er sich	auch wenn
nach dem	in dem
um die	bis zum
sich die	lässt sich

Demnach ist $\mu = 4,23^{-8}$. Das Wortpaar *berliner republik* kommt insgesamt 7 mal im Korpus vor; somit ist der Stichprobenmittelwert $\chi' = \frac{7}{255433} \approx 2,74^{-5}$. Die Standardabweichung kann auf $\chi' = 2,74^{-5}$ angenähert werden (siehe Manning and Schütze [2003], Seite 165). Somit ergibt sich der t-Wert durch

$$t = \frac{2,74^{-5} - 4,23^{-8}}{\sqrt{\frac{2,74^{-5}}{255433}}} \approx 2,6417.$$

Bei einem Signifikanzniveau $\alpha = 0.005$ (Konfidenzwert 2.576, s.o.) kann die Nullhypothese abgelehnt werden; das Wortpaar *berliner republik* ist eine Kollokation.

Es wurden die t-Werte für alle im Korpus vorkommenden Bigramme errechnet. Ich habe mich dazu des gesamten Korpus bestehend aus allen 180 der 10 Autoren bedient; der Umstand, dass hierbei auch die für die Klassifizierung vorbehaltene Testmenge von 60 Texten inbegriffen ist, stellt keine Schwierigkeit dar. Schließlich wird keinerlei Information (d.h. solche über die Verwendung bestimmter Bigramme durch die Autoren) in Gebrauch genommen, die später induziert werden soll. Außerdem sollen gerade diejenigen Kollokationen extrahiert werden, die in den Texten der Autoren auch tatsächlich vorkommen. Abbildung 1 zeigt die 50 Bigramme mit den höchsten t-Werten. Wie unschwer zu erkennen ist, finden sich unter diesen Bigrammen keine, die man als Kollokation bezeichnen würde. Allenfalls *zum beispiel* könnte eventuell man als semantisch nicht rein kompositionell betrachten.

Vermutlich lässt sich dieses Ergebnis auf die relativ kleine Größe des Trainingskorpus zurückführen. Um es zu verbessern, habe ich einen Filtermechanismus eingebaut, der alle diejenigen Bigramme ignoriert, deren erstes oder zweites Wort aus weniger als 5 Zeichen besteht. So wird ein Großteil der Funktionswörter enthaltenden Bigramme außer Acht gelassen und gleichzeitig die Implementierung einer Stoppwortliste umgangen. Abbildung 2 zeigt die 50 Bigramme mit den höchsten t-Werten nach der Filterung.

Neben vielen Personennamen finden sich in dieser Liste andere eindeutige Kollokationen, z.B. Ländernamen wie *vereinigten staaten*, Zeitperioden wie *achtziger jahre* oder Firmennamen wie *france télécó*. Anzumerken sei

Abbildung 2: Die 50 Bigramme mit den höchsten t-Werten nach dem Filtern von Stopwörtern

vereinigten staaten	boris jelzin
immer wieder	wolfgang schäuble
nicht einmal	schon immer
gerhard schröder	einen neuen
joschka fischer	seine partei
nicht zuletzt	france télécom
wladimir putin	eines tages
achtziger jahre	statt dessen
nicht gerade	einen namen
angela merkel	durch einen
milliarden dollar	rudolf scharping
vergangenen jahres	diesen tagen
millionen dollar	neunziger jahren
alles andere	prozent aller
telecom italia	heute nicht
neunziger jahre	nicht immer
nichts anderes	heute schon
schon heute	schon einmal
nicht alles	unter anderem
wenigen jahren	vielleicht sogar
diese weise	einen anderen
silicon valley	nicht genug
trotz aller	einer neuen
nicht länger	günter elsbett
einem anderen	antje vollmer

an dieser Stelle, dass zwar alle diese Bigramme einen t-Wert größer als der kritische Wert 2.576 bei $\alpha = 0.005$ besitzen, das Signifikanzniveau bei meinen Experimenten jedoch keine wichtige Rolle spielt; die t-Werte dienen vor allem der Möglichkeit, Bigramme nach ihrer Wahrscheinlichkeit, eine Kollokation darzustellen, sortieren und diejenigen Bigramme für das Training des Klassifizierers auswählen zu können, deren t-Werte am höchsten sind.

An dieser Stelle soll noch einmal explizit darauf hingewiesen werden, dass die hier eingesetzte Methode neben der Extraktion vieler Nicht-Kollokationen noch eine zweite Schwäche besitzt. Denn durch die Verwendung ausschließlich direkt benachbarter Wörtern sowie durch die Beschränkung auf Bigramme werden all solche Kollokationen übergangen, die aus mehr als zwei Wörtern bestehen oder eben durch andere Satzteile unterbrochen werden können. Abgesehen davon decken sich die Ergebnisse gut mit den Erwartungen von einer Implementation des t-Tests nach Manning and Schütze [2003] und den dort vorgestellten Resultaten.

3.3.2 Klassifikation

Zwei Dinge scheinen in Bezug auf die gefundenen Kollokationen in einem Text von Bedeutung zu sein: zum einen die Häufigkeit, mit der ein Autor eine bestimmte Kollokation benutzt. Es scheint aber gleichzeitig sinnvoll mit einzubeziehen, mit welcher Wahrscheinlichkeit es sich bei einem Wortpaar denn überhaupt um eine Kollokation handelt.

Zunächst habe ich die 100 Bigramme mit den höchsten t-Werten nach Filterung generiert. Als Merkmale für den von TiMBL implementierten maschinellen Lernalgorithmus wurde für jedes Bigramm im jeweiligen Trainingstext das Produkt aus t-Wert und relativer Häufigkeit berechnet. Algorithmus 2 zeigt den entsprechenden Pseudo-Code.

Algorithm 2 Berechnung der Merkmale mit Kollokationen

```
function getBigram_tScoreMap(trainingdata):
    bigramFreqDist = getBigramFreqDist(trainingdata)
    bigram_tScoreMap
    foreach bigram in bigramFreqDist:
        bigram_tScoreMap[bigram] = calculateTScore(bigram)
    return bigram_tScoreMap

collocations = getBigram_tScoreMap(trainingdata)
collocations.filterStopWords()
collocations.sort()
collocations = getFirst100Collocations(collocations)

foreach author:
    foreach text:
        bigramFreqDist = getBigramFreqDist(text)
        for c in collocations:
            value = bigramFreqDist.freq(c[0]) * c[1]
            writeToFeatVec(value + ",")
        writeToFeatVec(author)
```

3.3.3 Ergebnisse

Die Ergebnisse der beschriebenen Methode sind bescheiden (siehe Abschnitt 3.3.4): Die Akkuratheit beträgt rund 0,167 (10 von 60 Texten wurden korrekt klassifiziert). Tabelle 3 zeigt die entsprechende Konfusionsmatrix.

3.3.4 Fehleranalyse

Wieder fällt die ungleichmäßige Zuordnung der klassifizierten Texte zu Autoren auf. Allein 52% der falsch klassifizierten Texte (26 von 50 Texte) wurden Autor 1 zugewiesen. Dieses Ergebnis lässt sich leicht nachvollziehen, wenn man sich die Verteilung der Kollokationen über die Texte der einzelnen Autoren ansieht. So kommen in den Trainingstexten von Autor 1 lediglich 20 der 100 als Kollokationen veranschlagten Bigramme überhaupt vor (siehe Tabelle 4). In den entsprechenden Texten der Autoren 3, 4, 5 und 7 hingegen finden sich jeweils 45, 47, 47 bzw. 46 solcher Bigramme. Die Vermutung liegt nahe, dass der von TiMBL implementierte Trainingsalgorithmus Bigramme in Bezug auf einen bestimmten Autoren höher gewichtet, wenn dieser insgesamt nur wenige solche in seinen Texten benutzt.⁷

3.4 Zusätzliche Experimente

Zum weiteren Vergleich und ohne großen Mehraufwand ließen sich noch einige weitere Experimente umsetzen. Diese sollen im Folgenden knapp vorgestellt werden.

⁷Eine noch genauere Analyse wäre möglich, wenn man sich anschaut, welche Kollokationen von welchen Autoren wie oft eingesetzt wurden, eventuell sogar in Bezug auf die Verteilung in den 12 Trainingstexten. Dies liefe im Endeffekt auf eine Analyse des maschinellen Lernalgorithmus hinaus, worauf im Rahmen dieser Hausarbeit aus Platz- und Zeitgründen verzichtet werden muss.

Tabelle 3: Konfusionsmatrix der Kollokationsmethode

echte Klasse →	a10	a9	a8	a7	a6	a5	a4	a3	a2	a1	Fehler
a10	1	0	1	0	0	0	0	0	1	3	0,83
a9	0	0	0	0	0	0	0	0	0	6	1
a8	1	0	2	0	0	0	2	0	1	0	0,667
a7	2	1	0	0	0	1	0	0	0	2	1
a6	0	0	0	0	0	0	0	0	0	6	1
a5	0	0	1	0	1	0	0	0	0	4	1
a4	1	2	0	0	1	0	0	0	0	2	1
a3	0	1	0	0	1	0	0	0	3	1	1
a2	0	0	2	0	0	0	0	0	2	2	0,667
a1	0	0	0	0	1	0	0	0	0	5	0,167
falsch zugeordnete Texte	4	4	4	0	4	1	2	0	5	26	
	durchschnittl. Fehler:										0,833

Tabelle 4: Kollokationen pro Autor

Autor	Trainingsmenge	Testmenge
a1	20	10
a2	32	31
a3	45	31
a4	47	33
a5	47	31
a6	24	20
a7	46	22
a8	33	28
a9	36	14
a10	38	26

3.4.1 Kollokationen ohne vorheriges Filtern

Für oder wider Erwarten schneidet die Klassifikation anonymer Texte besser ab, wenn diejenigen Bigramme als Merkmale verwendet werden, die den höchsten t-Wert besitzen, jedoch ohne, dass Stopwörter enthaltende Bigramme vorher ausgefiltert wurden (siehe Tabelle 1). Merkmalsberechnung, Training und Klassifikation verlaufen abgesehen davon analog zum in Abschnitt 3.3.2 beschriebenen Vorgehen. Wohlgermerkt handelt es sich dabei nicht um Kollokationen im eigentlichen Sinn. Nichtsdestotrotz beträgt die Akkuratheit der Methode 30%, was einer richtigen Klassifikation von 18 aus 60 Texten entspricht.

3.4.2 Bigramme ohne t-Wert-Berechnung

Die in Abschnitt 3.4.1 vorgestellten Ergebnisse lassen vermuten, dass Bigramme ohne Betrachtung ihrer Kollokations- bzw. t-Wertes bei der Autorenerkennung erfolgreich sein könnten. Um dieser These nachzugehen, habe ich unter Verwendung der in NLTK 0.94 [Bird, 2006] zu Verfügung gestellten Funktionalität ein einfaches Bigramm-Modell des Korpus generiert, in welchem die Wahrscheinlichkeit jedes Bigramms mittels Maximum-Likelihood-Methode geschätzt wird. Die für TiMBL erstellten Merkmalsvektoren stellen die 100 häufigsten Bigramme dar, deren relative Frequenzen in den jeweiligen Texten der Trainingsmenge die Werte sind.

Die Methode klassifiziert 19 von 60 Texten richtig (Akkuratheit 0.317) und ist damit minimal besser als die im vorherigen Abschnitt 3.4.1 beschriebene. Diese Herangehensweise kann als äquivalent zur Baseline gesehen werden, die Bigramme anstelle von Unigrammen einsetzt. Insofern könnte man die Performanz, die ja nur ca. halb so gut wie die der Baseline ist, überraschend finden. De facto ist die Zerstreuung von Bigrammen jedoch wesentlich größer, d.h. es gibt wesentlich mehr Merkmale, deren Werte 0 betragen, was die niedrige Akkuratheit erklären würde.

3.4.3 Kombination von Kollokationen und Unigrammen

Die folgenden Ansätze kombinieren jeweils zwei der bisher vorgestellten Methoden. Die Anzahl der für das maschinelle Lernen benutzten Merkmale verdoppelt sich dabei jeweils auf 200.

Als erstes habe ich die Baseline und die Methode mit (gefilterten) Kollokationen kombiniert. Dabei fiel die Performanz der Baseline auf eine Akkuratheit von 0.47 (28 von 60 Texten korrekt klassifiziert). Dies entspricht einer Akkuratheit von weniger als dem Mittel der beiden Methoden. Wieder lässt sich dieses Ergebnis auf die starke Zerstreuung der Kollokationen zurückführen, die zusätzliches Rauschen in die Trainingsdaten einführt.

3.4.4 Kombination von ungefilterten Kollokationen und Unigrammen

Die Kombination von Baseline und ungefilterten Kollokationen (Abschnitt 3.4.1) bringt eine Verbesserung der Performanz auf 60% Akkuratheit (36 von 60 Texten korrekt klassifiziert). Dies liegt immer noch unter der Baseline-Performanz (63% Akkuratheit), jedoch nur sehr knapp.

3.4.5 Kombination von Bigrammen und Unigrammen

Wider Erwarten liegt die Performanz der Kombination der Bigramm-Methode (Abschnitt 3.4.2) mit der Baseline unter der im vorangegangenen Abschnitt 3.4.4 skizzierten Kombination von ungefilterten Kollokationen und Unigrammen. Dies deutet meines Erachtens darauf hin, dass der Unterschied zwischen der Bigramme-Methode und dem Ansatz mit ungefilterten Kollokationen zu vernachlässigen sein dürfte.

4 Fazit

In dieser Arbeit habe ich einen Ansatz zur Autorenerkennung vorgestellt, der Texte anhand der in ihnen vorkommenden Kollokationen klassifiziert. Es handelt sich dabei um den Versuch, die Bedeutungsebene bei der Bewältigung der Aufgabe mit in Augenschein zu nehmen. Dazu wurde das Korpus in ein Bigramm-Modell transformiert. Für jedes Bigramm wurde der t-Wert berechnet, der es ermöglicht, Bigramme nach ihrer Wahrscheinlichkeit zu ordnen, dass sie Kollokationen seien. Die Ergebnisse der Methode wurden vorgestellt und ausgewertet. Der Ansatz wurde einer Baseline gegenübergestellt, welche die relative Häufigkeit einzelner Wörter als Merkmale verwendet. Ein maschineller Lernalgorithmus stellte dabei die benötigte Klassifikationsfunktionalität zur Verfügung.

Wie bereits angedeutet, sind die Ergebnisse meines Ansatzes eher ernüchternd. Mehrere Gründe scheinen dabei eine Rolle zu spielen, einige habe ich bereits in den Fehleranalysen diskutiert. Zunächst einmal scheint es aber notwendig, die eigentliche Idee zu hinterfragen. Dass die Bigramme, die für die Klassifizierung verwendet wurden, in bei weitem nicht allen Fällen tatsächliche Kollokationen sind, ist das Resultat der Extraktionsmethode, die einen Hypothesentest verwendet (Abschnitte 2 sowie 3.3.1). Zweitens ist festzustellen, dass die Datenmenge vermutlich einfach nicht ausreicht, um befriedigende Ergebnisse zu bekommen. Gerade beim Übergang von Unigrammen zu Bigrammen wären entsprechend mehr Daten nötig, um vergleichbare Ergebnisse zu erreichen. Tatsächlich ist aber auch die ursprüngliche Zielsetzung der Verwertung der Bedeutungsebene nicht erfüllt, denn Kollokationen sind zwar ein Phänomen der lexikalischen Semantik, dennoch bewegt sich mein Klassifikationsansatz grundsätzlich auf der Ebene der Oberflächenformen. Denn wo ist der Unterschied zum traditionellen Ansatz, bestimmte, potentiell stilistisch bedeutsame, Wörter zur Klassifizierung heranzuziehen? Ich muss gestehen, dass mir keiner einfällt.

Des Weiteren wurde bisher nicht klar, ob Kollokationen tatsächlich den Stil eines Autors kennzeichnen. Beschränkt man sich auf reine Idiome, könnte dies so sein, müsste jedoch ebenfalls überprüft werden. Ein weiterer Blick auf Tabelle 2 legt jedoch die Vermutung nahe, dass durch Kollokationen, in diesem Fall ja viele Personennamen etc., eher das Thema des Textes beschrieben wird. Insofern sollte eine Folgestudie versuchen, eine Methode zur Extraktion reiner Idiome zu implementieren, welche, so weit es geht, einer theoretischen Motivation des Begriffs standhält. Zweitens sollte die Extraktion der Merkmalsvektoren so gestaltet werden, dass die Merkmalswerte weniger stark gestreut sind und eine bessere Gewichtung zulassen.

Literatur

- S. Bird. NLTK: the natural language toolkit. *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72, 2006.
- C.S. Brinegar. Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship. *Journal of the American Statistical Association*, 58(301):85–96, 1963.
- J.F. Burrows. Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*, 2(2):61–70, 1987.
- D. A. Cruse. *Lexical Semantics*, chapter 2.9. Cambridge Textbooks in Linguistics. Cambridge University Press, 1986.
- W. Daelemans and A. van den Bosch. *Memory-Based Language Processing*. Studies in Natural Language Processing. Cambridge University Press, 2005.
- W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. *TiMBL: Tilburg Memory-Based Learner*. http://ilk.uvt.nl/downloads/pub/papers/Timbl_6.1_Manual.pdf, December 2007. Version 6.1.

- J. R. Firth. Modes of meaning. *Papers in Linguistics 1934-1951*, 1957.
- W. Fucks. On Mathematical Analysis of Style. *Biometrika*, 39(1/2):122–129, 1952.
- D. Holmes. Authorship Attribution. *Computers and the Humanities*, 28(2):87–106, 1994.
- C.D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*, chapter 5. The MIT Press, 2003.
- J. Olsson. *Forensic Linguistics*. Continuum, 2008.
- HS Sichel. Word frequency distributions and type-token characteristics. *Mathematical Scientist*, 11(1):45–72, 1986.
- E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-Based Authorship Attribution Without Lexical Measures. *Computers and the Humanities*, 35(2):193–214, 2001.