Michael Tschuggnall

Automatisierte Plagiatserkennung in Textdokumenten: Was der Schreibstil eines Autors über die Echtheit verrät

2017

https://doi.org/10.25969/mediarep/1629

Veröffentlichungsversion / published version
Sammelbandbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
Tschuggnall, Michael: Automatisierte Plagiatserkennung in Textdokumenten: Was der Schreibstil eines Autors über die Echtheit verrät. In: Sandra Mauler, Heike Ortner, Ulrike Pfeiffenberger (Hg.): *Medien und Glaubwürdigkeit. Interdisziplinäre Perspektiven auf neue Herausforderungen im medialen Diskurs.* Innsbruck: Innsbruck University Press 2017 (Medien – Wissen – Bildung), S. 131–140. DOI: https://doi.org/10.25969/mediarep/1629.

Automatisierte Plagiatserkennung in Textdokumenten: Was der Schreibstil eines Autors über die Echtheit verrät

Michael Tschuggnall

Zusammenfassung

Aktuelle Ereignisse und die daraus resultierenden öffentlichen Diskussionen zeigen, dass Plagiate, und vor allem die Erkennung dieser, ein durchaus wichtiges Thema darstellen. Durch die hohe und ständig steigende Anzahl an frei verfügbaren Textdokumenten über literarische Datenbanken oder umfassende Online-Sammlungen wie z.B. Wikipedia wird es zudem immer leichter, Quellen für mögliche Plagiate zu finden, während es auf der anderen Seite für automatische Erkennungstools aufgrund der großen Datenmengen immer schwieriger wird, diesen entgegenzuwirken. Dieser Beitrag widmet sich der Frage, ob und wie sicher Verdachtsfälle über eine reine Analyse des Schreibstils, d.h. ohne Vergleiche mit existierenden Texten, erkannt werden können. Im Speziellen kann die unbewusste Verwendung von Grammatik zur Identifizierung von Autoren verwendet werden. Aktuelle Forschungsergebnisse zeigen, dass derartige Analysen erfolgreich für die Plagiatserkennung, aber auch für verwandte Gebiete wie z.B. die Autorenzuordnung oder Autorprofilerstellung eingesetzt werden können.

Einleitung

Durch die Entwicklungen im Bereich der elektronischen Datenverarbeitung und der Zugänglichkeit des World Wide Web steigt die Anzahl der öffentlich zugänglichen Textdokumente täglich. Neben Online-Bibliotheken wie z.B. Project Gutenberg, die Millionen von elektronischen Büchern zum freien Download anbieten, wird Text auch massiv über E-Mails, SMS oder Soziale Medien wie Facebook oder Twitter verbreitet. Im Gegensatz zu Letzteren, wo der Inhalt und die Autoren meist leicht klassifizierbar sind, stellt die unsachgemäße Wiederverwendung von Textfragmenten vor allem im akademischen Bereich ein ernsthaftes Problem dar. Die Erkennung solcher Plagiatsfälle kann mit recht einfachen Mitteln erfolgen, wenn ganze Textbausteine ohne oder mit nur geringfügiger Veränderung aus öffentlich zugänglichen und populären Quellen wie z.B. Wikipedia übernommen wurden. Andererseits ist es bereits wesentlich schwieriger, Plagiate zu erkennen, wenn der Originaltext stark umstrukturiert wurde, oder auch wenn die Quelle gar nicht (elektronisch) verfügbar ist. Vor allem in letzteren Fällen ist eine dokumentinterne Analyse des Schreibstils unvermeidbar. Während es für menschliche Leser oft leicht ist, Änderungen des Schreibstils zu identifizieren, ist dies für computerbasierte Algorithmen deutlich schwerer. Beispielsweise erkennen Betreuer von wissenschaftlichen Arbeiten Plagiate recht häufig daran, dass gewisse Sätze oder Absätze *anders* sind und nicht ins Gesamtbild passen. Eine Internetsuche nach dem entsprechenden Textfragment liefert dann oft rasch die Quelle.

Die Herangehensweise im obigen Beispiel verwendet dabei beide Möglichkeiten zur Plagiatserkennung: externe und intrinsische Erkennung. Bei einer externen Untersuchung wird ein Text nach gewissen Kriterien in Fragmente unterteilt, und jedes dieser Fragmente wird mit möglichen Quellen verglichen. Diese Quellen bestehen meist aus einer Kombination von großen Dokumentdatenbanken, die über die Zeit gesammelt und gespeichert worden sind, als auch aus über direkte Onlinesuche in verfügbaren Dokumenten. Die anschließend durchgeführten Vergleiche zwischen Text und Quelle werden dabei mithilfe von Algorithmen durchgeführt, die gegenüber kleineren Änderungen oder Fehlern robust sind, also z.B. mit Fehlern wie „Plagat“ anstatt „Plagiat“ umgehen können. Wird eine Übereinstimmung gefunden, so wird mit der externen Methode also auch immer automatisch die Quelle identifiziert.

Dies ist mit einer intrinsischen Analyse nicht möglich, da hier mögliche Plagiate ausschließlich dokumentintern durch gefundene Stiländerungen identifiziert werden (vgl. Potthast et al. 2011 sowie Stein et al. 2011). Oft verwendete Merkmale zum Erkennen von plötzlichen Stiländerungen in der intrinsischen Plagiatserkennung inkludieren z.B. die Art der Verwendung von Vokabular, die (durchschnittliche) Satzlänge oder die Komplexität der verwendeten Grammatik. Zusätzlich werden oft noch viele weitere lexikalische, syntaktische oder semantische Kenngrößen herangezogen, um Texte zu untersuchen. Unterscheidet sich der Stil in einem gewissen Textabschnitt signifikant vom Stil des gesamten Dokuments, so kann dies auf ein Plagiat hindeuten. Anders ausgedrückt kann also eine rein intrinsische Analyse nie hundertprozentige Sicherheit über das Vorkommen von Plagiaten liefern, sondern nur gut begründete und statistisch belegte Hinweise aufzeigen. Vollständige Gewissheit kann aber u.U. durch eine anschließende externe Suche über die relevanten Textstellen erlangt werden.

Im Folgenden wird ein Überblick über die in Tschuggnall (2014) entwickelten Methoden zur intrinsischen Plagiatserkennung gegeben, die insbesondere die verwendete Grammatik von Autoren untersuchen, um Stilbrüche zu erkennen. Anschließend wird aufgezeigt, welche zusätzlichen Anwendungsgebiete sich durch eine algorithmische Stilerkennung ergeben, darunter z.B. eine automatisierte Schriftstück-Autor-Zuordnung, das automatisierte Erstellen von Autorprofilen oder auch die Analyse von Bibeltexten.

Intrinsische Plagiatserkennung

Wie bereits erwähnt liefert die intrinsische Plagiatserkennung nur mögliche Verdachtsfälle, ist aber im Vergleich zur externen Erkennung weniger akkurat, da sie die Quelle im Gegensatz zu letzterer nicht liefert. Trotzdem werden intrinsische Methoden aus mehreren Gründen eingesetzt. Beispielsweise können Plagiate mit externen Algorithmen niemals gefunden werden, wenn sich die Quellen nicht in der Datenbank befinden oder nicht digital verfügbar sind (z.B. ältere Bücher oder akademische Abschlussarbeiten). Zusätzlich haben externe Erkenner oft Probleme mit zu stark modifizierten Textstellen, wenn etwa beispielsweise Wörter durch Synonyme ersetzt werden. Schlussendlich können intrinsische Ansätze vorgeschalten werden, um die benötigte Menge an Textvergleichen für externe Erkennungsmethoden signifikant zu reduzieren.

Der Stil von Autoren

Um den Schreibstil von Autoren zu quantifizieren werden verschiedenste Kennzahlen aus einem Text extrahiert. Häufig herangezogene Metriken sind dabei bevorzugtes Vokabular (vgl. Oberreuter et al. 2011), Phrasen, Wörter oder Buchstaben, verwendete Satz- und Sonderzeichen, Emoticons, Schriftfarben oder auch vorhandene Rechtschreib-, Grammatik- sowie Tippfehler (vgl. Koppel et al. 2003). Auch die verwendeten Grammatikstrukturen können systematisch analysiert werden, um zwischen verschiedenen Autoren zu unterscheiden. Diese ist insbesondere sehr hilfreich, da sie unabhängig von den konkret verwendeten Wörtern in den meisten Fällen unbewusst eingesetzt wird, um Sätze zu bilden. D.h. selbst bei bewusstem Austausch von Wörtern oder gar Manipulation von Teilsätzen bleibt die grundsätzliche Grammatikstruktur bestehen.

Jeder geschriebene Satz folgt den Grammatikregeln der verwendeten Sprache und kann durch einen sog. Grammatikbaum visualisiert werden. Dieser spiegelt den Aufbau der einzelnen Satzkomponenten wider und zeigt so implizit die verwendeten Regeln. Im Allgemeinen gibt es in einer Sprache sehr viele Möglichkeiten, Sätze syntaktisch zu formulieren, ohne dabei die Bedeutung zu verändern. Beispielsweise kann der englische Satz

> *"The strongest rain ever recorded in India shut down the financial hub of Mumbai, officials said today."* (S_1)

auch formuliert werden als

> *"Today, officials said that the strongest Indian rain which was ever recorded forced Mumbai's financial hub to shut down."* (S_2)

Die beiden Sätze sind semantisch äquivalent, unterscheiden sich aber signifikant in ihrer Syntax. Die Grammatikbäume, welche sich durch die Verwendung der entsprechenden Strukturregeln der englischen Sprache ergeben, sind in Abbildung 1 dargestellt. Die Knoten bezeichnen dabei sog. *Part-of-Speech (POS)-Tags* und klassifizieren so z.B. Verben (VB), Adjektive (JJ), Nomenphrasen (NP) oder Adverbphrasen (ADVP).

Der Plag-Inn-Algorithmus

Die Grundidee des in Tschuggnall & Specht (2013) entwickelten Algorithmus „Plag-Inn" besteht nun darin, etwaige Differenzen in den Grammatikbäumen wie in Abbildung 1 dargestellt zu quantifizieren, um so Irregularitäten in der verwendeten Syntax zu finden. Konkret besteht der Algorithmus aus den folgenden Schritten:

1. Bereinigung des Textes und Aufteilung in einzelne Sätze.
2. Berechnung eines Grammatikbaums für jeden Satz.
3. Berechnung des Unterschieds zwischen den einzelnen Grammatikbäumen.
4. Identifizierung von signifikant unterschiedlichen Sätzen (Bäumen).
5. Berechnung des finalen Ergebnisses, d.h. Markierung von potentiellen Plagiaten.

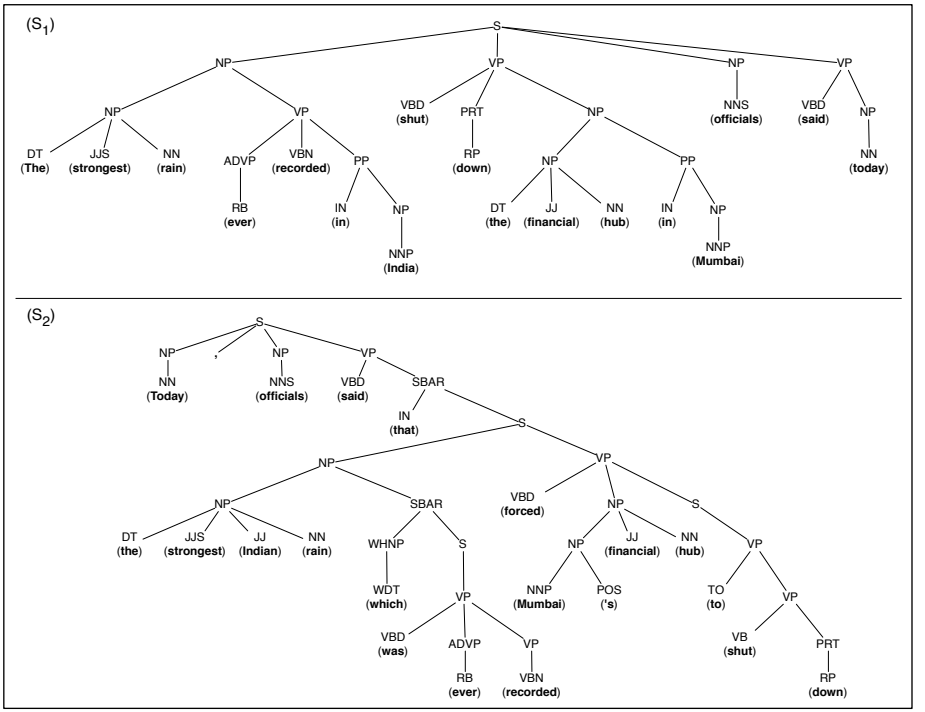

Abbildung 1: Grammatikbäume der Sätze S_1 und S_2.

Da Dokumente oft gescannt und über Texterkennung verarbeitet werden, entstehen Fehler bzw. unerwünschte Zeichen, die im ersten Schritt entfernt werden. Ebenfalls werden Formatierungen, kurze Überschriften oder Leerzeilen entfernt, da sie für die folgende grammatikalische Analyse nicht von Bedeutung sind. Nach diesen vorbereitenden Maßnahmen wird der Text nun in einzelne Sätze aufgeteilt. Dafür werden sog. „Sentence Boundary Disambiguation"-Algorithmen eingesetzt, die mit hoher Genauigkeit Satzgrenzen identifizieren (vgl. Palmer & Hearst 1997).

Für jeden Satz wird anschließend ein Grammatikbaum wie in Abbildung 1 berechnet, der als Grundlage für die folgenden Berechnungen dient. Liegen die Bäume aller Sätze vor, so wird für jedes Paar eine Distanz berechnet, d.h. ein Maß dafür, wie sehr sich die beiden Bäume unterscheiden. Diese Distanz wird mithilfe von sog. pq-Grammen (vgl. Augsten et al. 2010) bzw. der pq-Gramm-Distanz berechnet. Vereinfacht gesagt werden dabei alle möglichen grammatikalischen Teilstrukturen aus einem Baum nach gewissen Regeln extrahiert und anschließend mit den Teilstrukturen des zweiten Baums verglichen. Da der Plag-Inn-Algorithmus

ausschließlich die verwendete Grammatik überprüft, werden die konkret verwendeten Wörter, d.h. das Vokabular, ignoriert. Je mehr Teilstrukturen übereinstimmen, desto ähnlicher sind sich die Bäume. Die Distanz zwischen allen Bäumen, repräsentiert durch einen numerischen Wert, wird dann in einer Matrix festgehalten, die visualisiert in etwa wie in Abbildung 2 aussehen kann.

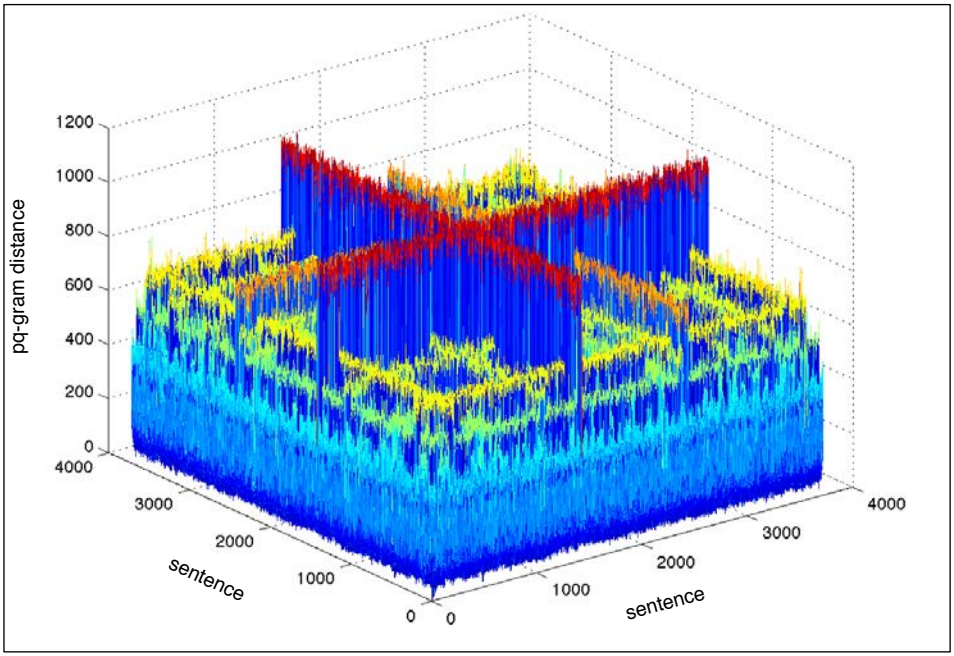

Abbildung 2: Grammatikbäume der Sätze S_1 und S_2.

Das Beispiel zeigt ein Dokument mit ca. 4.000 Sätzen, wobei x- und y-Achsen die Position der Sätze im Dokument repräsentieren. Die Höhe der Werte der z-Achse zeigen schlussendlich die Distanzen zwischen den Sätzen. Im Beispiel ist bereits mit freiem Auge ersichtlich, dass gewisse Sätze sich im Vergleich zu allen anderen deutlich mehr unterscheiden, grammatikalisch also einen deutlich anderen Stil aufweisen. Um dies auch algorithmisch zu quantifizieren, wird die durchschnittliche Distanz für jeden Satz berechnet und anschließend werden mit statistischen Standardverfahren Ausreißer ermittelt.

Die ermittelten Ausreißer, d.h. Sätze, die sich unter Betrachtung des grammatikalischen Schreibstils signifikant unterscheiden, werden als potenziell plagiiert gekennzeichnet und dienen schlussendlich als Ausgangspunkt für die finale Berechnung des Endergebnisses. Dafür

wurde ein Algorithmus entwickelt, der das Ergebnis „glättet“, also von Unsicherheiten so gut wie möglich befreit. Beispielsweise werden einzelne Sätze wieder als nicht plagiiert gekennzeichnet, wenn sie sich isoliert inmitten von nicht plagiierten Sätzen befinden. Umgekehrt werden Sätze auch erst im letzten Schritt als Plagiat markiert, wenn sich diese innerhalb eines größeren Plagiatbereichs befinden. Dies kann z.B. auftreten, wenn kurze Bindesätze kopiert wurden, die zu wenig grammatikalische Information enthalten, um alleinstehend als Ausreißer identifiziert zu werden. Deshalb wird in solchen Fällen auch immer der Kontext miteinbezogen.

Evaluationen auf einem speziell erzeugten Datenset zeigen, dass der Algorithmus eine Genauigkeit von über 35% erreicht, was einem hohen Wert für intrinsische Verfahren entspricht. Weiters konnte erforscht werden, dass der Algorithmus bei Normallängen-Dokumenten von ca. 100-200 Sätzen (etwa ein wissenschaftlicher Artikel) im Vergleich zu Buchlänge-Dokumenten noch deutlich an Zuverlässigkeit gewinnt und bis zu 50% Genauigkeit erzielt.

Varianten

Um die Genauigkeit weiter zu erhöhen, wurden noch zwei zusätzliche Varianten des Plag-Inn-Algorithmus entwickelt, die auf der originalen Idee aufbauen, sich aber in mehreren Details unterscheiden. Im *POS-Plag-Inn*-Ansatz wird auf die Auswertung von Grammatikbäumen verzichtet und stattdessen werden nur linearisierte, „flachgedrückte“ Folgen von Part-of-Speech-Tags verarbeitet. Zum Beispiel wird für den Satz *„This is a simple sentence“* kein Grammatikbaum, sondern nur die Folge „DT-VBZ-DT-JJ“ berechnet. Die Distanzen zwischen Sätzen werden dann mit diesen Folgen berechnet, was mithilfe von adaptierten Algorithmen aus der Genetik (Sequenz-Alinierung) durchgeführt wird.

Weiters wurde im *PQ-Plag-Inn*-Algorithmus eine Variante entwickelt, die nicht mehr einzelne Sätze miteinander vergleicht, sondern Profile aus pq-Grammen, d.h. grammatikalischen Teilstrukturen erstellt. Hier wird für das gesamte Dokument ein Profil erstellt, welches die meistverwendeten Teilstrukturen und deren Häufigkeiten enthält. Anschließend werden schrittweise einzelne Textabschnitte untersucht, deren Profile berechnet und mit dem Profil des gesamten Dokuments verglichen. Unterscheiden sich Textabschnitte signifikant, so werden diese als mögliches Plagiat eingestuft.

Beide Varianten wurden ebenfalls ausgiebig getestet und optimiert, und es zeigt sich, dass der POS-Ansatz in etwa dieselbe Performanz wie der Grundalgorithmus liefert, während die Variante mit den Profilen nochmal eine deutliche Verbesserung erzielen kann.

Weitere Anwendungsgebiete

Eine stilistische Analyse von Text hat neben der intrinsischen Plagiatserkennung noch viele weitere Einsatzgebiete, von einer automatischen Zuordnung von Texten zu Autoren über die Generierung von Empfehlungen bis hin zu klinischen Anwendungen. Im Folgenden wird ein kurzer Überblick gegeben, was konkret mit der bisher beschriebenen grammatikalischen Analyse entwickelt wurde.

Zuordnung von Autoren

Die Problemstellung der automatischen Autorenerkennung kann recht einfach formuliert werden: Weise einem Textdokument unbekannter Urheberschaft einen bekannten Autor zu, oder anders formuliert: Gegeben sei ein Textdokument – wer hat es geschrieben? Die Zahl der möglichen Autoren wird dabei meist so weit wie möglich eingeschränkt (auf z.B. drei oder maximal 20), und für jeden Kandidaten existieren verifizierte Schriftstücke, mit denen das unbekannte Dokument verglichen werden kann. Aktuelle Ansätze verwenden eine Reihe von Kenngrößen (oft mehr als 100), welche sehr häufig mit sog. Machine-Learning-Algorithmen verarbeitet werden (vgl. Stamatatos 2009).

In Tschuggnall & Specht (2014b) wurde der entwickelte Ansatz zur Plagiatserkennung so adaptiert, dass er für die Autorenerkennung verwendet werden kann. Aufgrund der Evaluationen zur Plagiatserkennung, die für den Profilansatz die besten Ergebnisse lieferten, wurde auch hier mit Profilen gearbeitet, die auf Teilstrukturen von Grammatikbäumen basieren. Die Zuordnung erfolgt anschließend so, dass selbstlernende Machine-Learning-Algorithmen mit bekannten Texten von Autoren „trainiert“ werden, sodass schlussendlich nach selbst abgeleiteten Regeln ein neues, unbekanntes Schriftstück einem dieser Autoren zugewiesen werden kann. Mit einer Genauigkeit von über 75-90% konnten sowohl Datensätze mit wenigen als auch mit mehreren Kandidaten zugewiesen werden, wobei einzelne Datensätze sogar eine Genauigkeit von 100% erreichten. Dies ist insbesondere deshalb ein ausgesprochen gutes Resultat, weil im Vergleich zu anderen Ansätzen hier nur eine einzige Kenngröße herangezogen wurde, und zwar die verwendete Grammatik der Autoren.

Ein verwandtes Problem wurde in Tschuggnall & Specht (2014a) behandelt. Dabei wurde versucht, Einzelbeiträge aus einem gemeinschaftlich geschriebenen Dokument zu filtern. Die Arbeitsweise ist hierbei sehr ähnlich zu den vorigen Ansätzen, d.h. es wurde wieder mit Grammatikbäumen, Profilen und Machine-Learning-Algorithmen gearbeitet. Da es allerdings möglich sein sollte, auch ohne vorher bekannte Proben von den Autoren Einzelbeiträge zu finden, musste auf andere Algorithmen zurückgegriffen werden, die imstande sind, ohne jedes Vorwissen nach grammatikalischen Ähnlichkeiten Gruppierungen zu erstellen. Eine Evaluation wurde auf verschiedenen Testdatensätzen durchgeführt und erreichte – gemittelt über alle Datensätze – eine Genauigkeit von etwa 63%.

Erstellen von Autorenprofilen

In aktuellen *Profiling*-Ansätzen wird versucht, möglichst viel Information aus einem gegebenen Textstück zu extrahieren. Dies umfasst häufig das Alter und Geschlecht des Autors (vgl. Argamon et al. 2009), aber auch Daten wie den kulturellen Hintergrund, den Ausbildungsgrad oder psychologische Einstufungen wie etwa Intro-/Extrovertiertheit (vgl. Noecker et al. 2013). Eine automatisierte Früherkennung von Depression aufgrund des sich ändernden Schreibstils wurde in Losada et al. (2016) prototypisch entwickelt und mit vielversprechenden Ergebnissen evaluiert.

In Tschuggnall & Specht (2014c) wurde versucht, sowohl das Geschlecht als auch das Alter (aufgeteilt in drei Altersgruppen) des Autors eines gegebenen Schriftstücks aufgrund der verwendeten Grammatik automatisch zu erkennen. Dabei wurde ähnlich wie bei der Autorenerkennung wieder auf pq-Gramm-Profile zurückgegriffen, die wieder mit selbstlernenden Algorithmen verarbeitet wurden. Die Evaluation auf einem Testdatenset von mehreren tausend Web-Blogs ergab auch hier sehr gute Ergebnisse: Das Geschlecht konnte mit nahezu 70% Genauigkeit erkannt werden, und das Alter mit knapp über 60%. Auswertungen im Detail ergaben, dass die Grammatikanalyse ausgesprochen gut zwischen Personen im Alter von 10-20 und Personen von 20-30 unterscheiden kann, allerdings Probleme bei der Trennung zwischen Letzteren und Personen über 30 hat.

Bibelanalyse

Eine weitere interessante Anwendung findet die entwickelte Grammatikanalyse in der Literarkritik. In Tschuggnall et al. (2016) wurde versucht, Bibelstellen nach Autoren zu gruppieren sowie „Plagiate“ zu finden, d.h. Stellen innerhalb eines Verses zu finden, die einen signifikant anderen Schreibstil aufweisen. So wurden angepasste Algorithmen unter Verwendung des von Wolfgang Richter entwickelten Grammatikmodells auf althebräischen Text angewandt. Sowohl in der „Plagiaterkennung“ als auch in der Gruppierung von Autoren konnten theologisch als gesichert geltende Fakten algorithmisch belegt werden, teilweise mit Übereinstimmungsraten von 100%. Neben diesen Bestätigungen wurden auch noch nicht bekannte Unstimmigkeiten lokalisiert, die als Grundlage für künftige literarkritische Forschung dienen soll.

Zusammenfassung

Textueller Plagiarismus ist ein häufig auftretendes Problem in der modernen vernetzten Gesellschaft, vor allem durch die leichte Zugänglichkeit von Millionen von Textdokumenten. Als eine mögliche Gegenmaßnahme wurden in diesem Artikel intrinsische Plagiatserkennungsalgorithmen vorgestellt, die ausschließlich das zu prüfende Dokument untersuchen und keine externen Vergleiche durchführen. Die Grundidee, dass sich Autoren in der Verwendung ihrer Grammatik signifikant unterscheiden, diese aber meist unbewusst verwenden und somit ungewollte Fingerabdrücke hinterlassen, wurde systematisch verfolgt und in mehreren Varianten validiert. Evaluationen und Optimierungen liefern sehr gute Ergebnisse und deuten darauf hin, dass Plagiate tatsächlich durch Grammatikanalysen gefunden werden können. Weiter wurde gezeigt, dass dieselben Ideen auch bei verwandten Problemstellungen wie der automatischen Erkennung von Autoren, Extrahieren von Alter und Geschlecht oder der Analyse von alten Texten erfolgreich eingesetzt werden können.

Literatur

Augsten, Nikolaus; Böhlen, Michael; Gamper, Johann (2010): The pq-gram Distance Between Ordered Labeled Trees. In: ACM Transactions on Database Systems (TODS) 35 (1), S. 4.

Argamon, Shlomo; Koppel, Moshe; Pennebaker, James W. & Schler, Jonathan (2009): Automatically Profiling the Author of an Anonymous Text. Communications of the ACM 52 (2), S. 119-123.

Koppel, Moshe & Schler, Jonathan (2003): Exploiting Stylistic Idiosyncrasies for Authorship Attribution. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico, S. 72-80.

Losada, David E. & Crestani, Fabio (2016): A Test Collection for Research on Depression and Language Use. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, S. 28-39.

Noecker, John; Ryan, Michael & Juola, Patrick (2013): Psychological Profiling Through Textual Analysis. In: Literary and Linguistic Computing 28 (3), S. 382-387.

Oberreuter, Gabriel; L'Huillier, Gaston; Rios, Sebastian A. & Velasquez, Juan D. (2011): Approaches for Intrinsic and External Plagiarism Detection. In: Notebook Papers of the 5th Evaluation Lab on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN). Amsterdam, The Netherlands.

Palmer, David D. & Hearst, Marti A. (1997): Adaptive multilingual sentence boundary disambiguation. In: Computational Linguistics 23 (2), S. 241-267.

Potthast, Martin; Eiselt, Andreas; Barron-Cedeno, Alberto; Stein, Benno & Rosso, Paolo (2011): Overview of the 3rd International Competition on Plagiarism Detection. In: Note-

book Papers of the 5th Evaluation Lab on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN). Amsterdam, The Netherlands.

Stamatatos, Efstathios (2009): A Survey of Modern Authorship Attribution Methods. In: Journal of the American Society for Information Science and Technology 60 (3), S. 538-556.

Stein, Benno; Lipka, Nedim & Prettenhofer, Peter (2011): Intrinsic Plagiarism Analysis. In: Language Resources and Evaluation 45 (1), S. 63-82.

Tschuggnall, Michael & Specht, Günther (2013): Detecting Plagiarism in Text Documents Through Grammar-Analysis of Authors. In: Proceedings of the 15th Fachtagung des GI-Fachbereichs Datenbanksysteme für Business, Technologie und Web (BTW). LNI, GI, Magdeburg, Germany, S. 241-259.

Tschuggnall, Michael (2014): Intrinsic Plagiarism Detection and Author Analysis By Utilizing Grammar. Dissertation, Institut für Informatik, Universität Innsbruck.

Tschuggnall, Michael & Specht, Günther (2014a): Automatic Decomposition of Multi-Author Documents Using Grammar Analysis. In: Proceedings of the 26th GI-Workshop on Grundlagen von Datenbanken. CEUR-WS, Bozen, Italy.

Tschuggnall, Michael & Specht, Günther (2014b): Enhancing Authorship Attribution By Utilizing Syntax Tree Profiles. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), volume 2: Short Papers. Association for Computational Linguistics, Gothenburg, Sweden, S. 195-199.

Tschuggnall, Michael & Specht, Günther (2014c): What Grammar Tells About Gender and Age of Authors. In: Proceedings of the 4th International Conference on Advances in Information Mining and Management (IMMM). Paris, France, S. 30-35.

Tschuggnall, Michael; Specht, Günther & Riepl, Christian (2016): Algorithmisch unterstützte Literarkritik. In H. Rechenmacher (Hrsg.): Arbeiten zu Text und Sprache im Alten Testament, 100. Band (ATSAT 100), In Memoriam Wolfgang Richter. St. Ottilien, 2016.