Plagiatserkennung

von Textplagiaten im wissenschaftlichen Kontext

Was ist ein Plagiat?

Ein Plagiat ist eine „unbefugte Verwertung unter Anmaßung der Autorschaft' [und damit] eine Verletzung geistigen Eigentums" [1].
Verschiedene Formen des Plagiats im wissenschaftlichen Kontext sind (jeweils dann, wenn die Quelle nicht angegeben wird) nach [1]:

- Textplagiat (wörtliches Plagiat): wörtliche Übernahmen aus fremden Texten
- Ideenplagiat (auch: Strukturplagiat, paraphrasierendes Plagiat): fremde Gedankengänge in eigenen Worten wiedergeben
- Übersetzungsplagiat: Übersetzung von Textpassagen und Gedankengängen aus einem fremdsprachigen Werk
- Zitatsplagiat: Zitate, die aus Texten übernommen werden (Zitat vom Zitat)
- Imitationsplagiat: Übernahme von prägnanten Formulierungen oder sprachlichen Schöpfungen wie Metapher

Eine Sonderform ist „collusion", die unerlaubte Zusammenarbeit von Studierenden an einer Aufgabe, die eigentlich eigenständig bearbeitet werden muss. [2]

Wie werden Plagiate erkannt?

Reine Textplagiate sind relativ einfach zu erkennen, aber Plagiate werden oft verschleiert, zum Beispiel durch Löschung von Worten, Ersetzung von Worten durch Synonyme, Veränderung der Wortreihenfolge oder des Tempus oder Aktiv-Passiv-Wechsel [2], aber auch durch Einsetzen von Homoglyphen oder weißen Buchstaben statt Leerzeichen [3].
Algorithmen zur Erkennung von Plagiaten können in externe und intrinsische Methoden aufgeteilt werden [2].

EXTERNE PLAGIATSERKENNUNG: [2]

Bei der externen Plagiatserkennung wird der Text mit einem Korpus (Sammlung von Texten) verglichen. Der Korpus wird schon vorher kompiliert, je nach Methode auf unterschiedliche Weise. Es gibt zahlreiche Methoden:

- Analyse von einzelnen Wörtern
- Untersuchung von mehreren Wörtern, z.B. werden beim Ferret-System überlappende Sequenzen von drei Wörtern gebildet
- Syntaktische Prozessierungstechniken wie Bildung von Wortgruppen anhand von syntaktischer Analyse („depencendy parsing and chunking") ermöglichen auch die Erkennung von Funktionen von Wörtern und Satzbausteinen, um Paraphrasierungen und Umstellungen zu erkennen

Bei jeder Methode sollte das Ergebnis für jeden Vergleich von zwei Texten ein Wert auf einer Skala sein. Eine funktionierende Methode erreicht den Maximalwert der Skala, wenn zwei gleiche Texte verglichen werden, und einen Wert von etwas unter 1 % bei „Hintergrundrauschen" (plagiatsfreie Texte über ein ähnliches Thema). Bei hohen Werten muss dann manuell untersucht werden, ob tatsächlich ein Plagiat vorliegt. Es gibt zum Beispiel im Rahmen der PAN-Konferenz auch Vergleiche von verschiedenen Methoden, die auf den selben Korpus und Vergleichstext angewendet werden - auf diese Weise können die Methoden bewertet werden. [4]
Zur Veranschaulichung ein Beispiel für eine Methode von Chong [5] mit n-Grammen[1] aus Wörtern. Chong verwendet 3-Gramme, zur Erklärung werden 2-Gramme benutzt.
In einem Text des Korpus wird für jedes Wortpärchen die Wahrscheinlichkeit bestimmt, dass das erste Wort in Verbindung mit dem zweiten Wort auftritt. Kommt „red apple" zwei mal vor, „red" insgesamt acht mal, ist $P(red, apple) = 2/8$. Das Ergebnis der Textanalyse ist eine Liste von Wahrscheinlichkeiten für jedes Wortpaar.

[1] N-Gramme sind Teile eines Textes, die zum Beispiel aus n Wörtern oder n Buchstaben bestehen.

Ein Beispieltext für den Vergleich ist „I ate a red apple“. Dieser Text wird auch auf Wortpaare durchsucht, und die Wahrscheinlichkeiten für die gefundenen Wortpaare aus dem Korpustext werden zum „bigram score“ multipliziert (die Zahlenwerte sind Beispiele):
$bigramscore = P(I, eat) * P(eat, a) * P(a, red) * P(red, apple) = 1/6 * 5/8 * 7/9 * 2/8 = 70/3456$
Weil Texte verschieden lang sind, muss dieser Wert noch mit der Textlänge m zur „Perplexity“ normiert werden:
$Perplexity = 1/m * log_2(bigramscore) = 1/5 * log_2(70/3456) = -2{,}813$
Die Perplexities im Verhältnis zu jedem einzelnen Korpustext können dann auf der Skala einsortiert werden. Hohe Werte bedeuten eine hohe Übereinstimmung.
Statt benachbarten Wörtern kann man in dieser Methode zum Beispiel auch semantisch zusammenhängende Wörter einlesen.

Problematik von externen Methoden
Paraphrasierungen und vor allem Übersetzungen sind in externen Methoden schwer zu erkennen. Es gibt aber auch die grundsätzliche Schwierigkeit, dass nicht alle möglichen Plagiatsquellen elektronisch verfügbar sind.

INTRINSISCHE PLAGIATSERKENNUNG [6]:
Intrinsische Methoden erkennen Inkonsistenzen im Schreibstil des möglichen Plagiats. Die Voraussetzung dafür ist, dass der Text kein Komplettplagiat ist oder von einem Ghostwriter geschrieben wurde.
Der erste Schritt ist, den Text in Abschnitte einzuteilen, zwischen denen dann Stilbrüche erkannt werden können. Beispiele sind, in 200-Wort-Schritten Ketten aus den nächsten 1000 Wörtern zu bilden [7], aber auch Absätze, Kapitel oder thematische Abschnitte mit dafür geeigneten Methoden zu erkennen [6].
Der zweite Schritt ist die Anwendung von stylometrischen Messgrößen. Es gibt davon sehr viele, die in verschiedene Kategorien eingeordnet werden können:
- Buchstabenbasierte lexikalische Messgrößen, z.B. n-Gramm-Frequenz/-Verhältnis von Buchstaben
- Wortbasierte lexikalische Messgrößen, z.B. durchschnittliche Wortlänge oder durchschnittliche Anzahl von Silben pro Wort
- Syntaktische Messgrößen, z.B. Häufigkeit von Funktionswörtern oder n-Gramm-Frequenz/-Verhältnis von Satzbausteinen
- Strukturelle Messgrößen, z.B. Häufigkeit von Absätzen
- Kombinierte Messgrößen wie die Lesbarkeit (der Flesch-Lesbarkeitsindex berechnet sich z.B. aus der durchschnittlichen Satzlänge und der durchschnittlichen Wortlänge [8])

Abschließend kann dann zum Beispiel wieder die Häufigkeit von n-Grammen berechnet und auf einen vorher festgelegten Schwellwert für einen Stilwechsel untersucht werden [2].
Auch Übersetzungen können so gefunden werden, da sie geringeren Wortschatz haben, lesbarer und kürzer sind. Auch die Herkunftssprache lässt sich automatisch erkennen [2].

Problematik von internen Methoden
Logischerweise kann nicht direkt überprüft werden, von wo plagiiert wurde. Außerdem kann sich der Stil in einem Dokument auch einfach so schon mal ändern, zum Beispiel weil sich während einer mehrjährigen Dissertation die Stimmung der Autorin oder des Autors ändert [9].

Welche Software gibt es?
Turnitin, Copyscape, Urkund sind die drei Softwares, die von Weber-Wulff als „teilweise nützlich“ eingestuft wurden. Turnitin und Urkund: nutzen drei Suchbereiche: das Internet, veröffentlichte wissenschaftliche Arbeiten und bereits geprüfte Einreichungen (wie etwa studentische Arbeiten). Urkund prüft dabei auch Synonyme und Paraphrasierungen. [10,11] Vorgehensweisen oder exakte Algorithmen von den Softwaretools sind nicht veröffentlicht.

Wie gut funktioniert das in der Praxis?
Debora Weber-Wulff nennt Plagiatserkennungssoftware nach 15 Jahren Erfahrung damit, sie zu testen, „eine Krücke und ein Problem“. [12]

Warum eine Krücke?
Bei Redewendungen, langen Institutionsnamen und Quellenangaben werden Plagiate erkannt, obwohl es keine sind. Wenn die Originale nicht digitalisiert sind, die Plagiate Tippfehler enthalten, bei Übersetzungen und bei Zusammensetzungen aus vielen verschiedenen Quellen schlagen die Softwares nicht an. Außerdem erkennen verschiedene Softwares den selben Text als Plagiat, als teilweise plagiiert und als plagiatsfrei, und Passagen mit korrekten Literaturangaben werden als Plagiate bezeichnet. Manche Systeme wählen nur zufällig Vergleichstexte aus dem Korpus aus und geben deshalb unterschiedliche Ergebnisse bei mehrfacher Überprüfung aus. Dazu kommt, dass die Bedienbarkeit und Übersichtlichkeit oft sehr schlecht ist. [12]

Warum ein Problem?
Diejenigen, die die Texte lesen, stehen oft unter Zeitdruck. Sie nehmen deshalb den „Score", den die Software ausgibt, als Bewertungsmaßstab, ohne sich die Ergebnisse genau anzusehen, und verlassen sich auf ein System, ohne eine „zweite Meinung" von anderer Software einzuholen. Eigentlich müsste man aber jeden möglichen „Treffer" analysieren, bevor man Studierende „vorverurteilt". Auf der anderen Seite werden bei guten Ergebnissen offensichtliche Plagiate, die an Formatierungs- und Stilveränderungen leicht erkannt werden könnten, übersehen. Wenn die nötige Punktzahl dafür, dass eine Arbeit an der Uni oder bei einem Journal akzeptiert wird, sogar öffentlich ist, kann der Text paraphrasiert werden, bis die Punktzahl erreicht ist. [12]

Was ist ihr Fazit?
Ihre Empfehlung ist, dass erst einmal Stilveränderungen und ungewöhnliche Quellen von einem Menschen gefunden werden und mit Hilfe von Suchmaschinen überprüft werden. Nur bei Texten, die wie ein Plagiat wirken, aber so nicht überführt werden können, sollte die Software für die Markierung von möglichen Problemen verwendet werden. Die Kontrolle dieser Textstellen muss dann aber wieder ein Mensch übernehmen. [12]

Quellen

[1] Universität Duisburg-Essen (2019): *Plagiate (Definition)*. Online erhältlich unter https://www.uni-due.de/de/gute-wissenschaftliche-praxis/plagiate.php, zuletzt abgerufen am 17.06.2019.

[2] Oakes, Michael P. (2014): *Author Profiling and Related Applications*. In Mitkov, Ruslan (Hrsg.): The Oxford Handbook of Computational Linguistics 2nd edition. Oxford University Press.

[3] Weber-Wulff, Debora; Möller, Christopher; Touras, Jannis; Zincke, Elin (2013): *Plagiarism Detection Software Test 2013*. In HTW Berlin: Plagiats Portal. Online erhältlich unter: http://plagiat.htw-berlin.de/software-en/test2013/report-2013/, zuletzt abgerufen am 17.06.2019.

[4] Potthast, Martin; Gollub, Tim; Hagen, Matthias; Stein, Benno; Rosso, Paolo: *PAN @ CLEF 2015 - Plagiarism Detection*. Online erhältlich unter https://pan.webis.de/clef15/pan15-web/plagiarism-detection.html, zuletzt abgerufen am 17.06.2019.

[5] Chong, Miranda; Specia, Lucia; Mitkov, Ruslan (2010): *Using Natural Language Processing for Automatic Detection of Plagiarism.* Proceedings of the 4th International Plagiarism Conference (IPC-2010). Online erhältlich unter: https://www.researchgate.net/profile/Lucia_Specia/publication/242783426_Using_Natural_Language_Processing_for_Automatic_Detection_of_Plagiarism/links/56179c2908ae0224ebce9956.pdf, zuletzt abgerufen am 17.06.2019.

[6] Stein, Benno; Lipka, Nedim; Prettenhofer, Peter (2011): *Intrinsic plagiarism analysis*. Language Resources and Evaluation, 45(1), S. 63-82.

[7] Stamatatos, Efstathios (2009): *Intrinsic Plagiarism Detection Using Character n-gram Profiles*. threshold 2(1,500).

[8] DuBay, William H. (2004): *The Principles of Readability.* Online erhältlich unter: https://files.eric.ed.gov/fulltext/ED490073.pdf, zuletzt abgerufen am 17.06.2019.

[9] Infinitesimalia, Sophia A. A.(2011): *Plagiatserken nung - ein steiniger Weg für die Computerlinguistik*. In: Frankfurter Allgemeine Blogs: Deus ex Machina. Online erhältlich unter: https://blogs.faz.net/deus/2011/08/04/plagiatserkennung-ein-steiniger-weg-fuer-die-computerlinguistik-478/, zuletzt abgerufen am 17.06.2019.

[10] Urkund: *About Urkund - The solution to plagiarism problems*. Online erhältlich unter: https://www.urkund.com/de/about-urkund/, zuletzt abgerufen am 17.06.2019.

[11] Turnitin: *Database Content | Turnitin*. Online erhältlich unter: https://www.turnitin.com/about/content, zuletzt abgerufen am 17.06.2019.

[12] Weber-Wulff, Debora (2019): *Plagiarism detectors are a crutch, and a problem*. Nature, 567, S. 435. Online erhältlich unter: https://www.nature.com/magazine-assets/d41586-019-00893-5/d41586-019-00893-5.pdf, zuletzt abgerufen am 17.06.2019.