Zum Beispiel Plagiatur: Sprachtechnologie für den Einsatz in der Hochschule

Aljoscha Burchardt, Prof. Iryna Gurevych, UKP Lab

elc-Vortragsreihe „E-Learning“, 02/09

TECHNISCHE UNIVERSITÄT DARMSTADT

Ubiquitous Knowledge Processing

Zum Vortragenden

TECHNISCHE UNIVERSITÄT DARMSTADT

Aljoscha Burchardt

- Computerlinguist
- Wissenschaftlicher Koordinator des Forschungsschwerpunktes E-Learning an der TUD

e-learning center of research excellence

- Area-Head am UKP Lab (*Ubiquitous Knowledge Processing*)
- Sprachtechnologie/NLP (Natural Language Processing) für E-Learning
  - Semantisches Informationsmanagement (Suche, Zusammenfassung, Ablage)
  - Automatisches Qualitätssicherung
  - Schwerpunkt benutzergenerierte Inhalte (Blogs, Foren, Wikis)
  - …

Was ist Plagiatur?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Im Allgemeinen:

  Das illegitime und illegale Veröffentlichen oder Verwenden von geistigem Eigentum oder Erkenntnissen, die sich eine andere Person erarbeitet oder geschaffen hat mit dem Ziel sich darüber einen persönlichen Vorteil zu verschaffen.

  (http://de.wiktionary.org/wiki/Plagiat)

- Hauptmerkmal: Fehlende Kenntlichmachung

Betroffene Trägermedien

- Musikstücke
- Texte
- Grafiken
- Bilder
- …

- Im Vortrag: Beschränkung auf Textmaterial

Plagiatur an der Hochschule

TECHNISCHE UNIVERSITÄT DARMSTADT

- Zwei verbreitete Arten von Plagiatur unter Studierenden
  - Intra-corpal plagiarism
    - Abschreiben bei Mitstudierenden
    - Kollusion (hier: unerlaubte Gruppenarbeit)
  - Web-based plagiarism
    - Abschreiben aus einer online-Quelle (Buch, Webseite)

(Culwin and Lancaster 2001)

- „Web 2.0“-Mentalität: *Find-Remix-Share*

(Sattler 2007)

Plagiatur an der Hochschule (Lehrende/Forscher)

TECHNISCHE UNIVERSITÄT DARMSTADT

- In Lehrmaterialien: Folien / Kurs-Reader / etc.

- Self-plagiarism
- Stillschweigende Übernahme von Ergebnissen für die eigene Arbeit (von Doktoranden, Diplomanden, etc.)

(http://www.spiegel.de/unispiegel/jobundberuf/0,1518,207062,00.html)

- Peer-Reviews (Projektanträge, Konferenzpapiere)

(http://de.wikipedia.org/wiki/Plagiat#Plagiate_in_Hochschule_und_Schule)

- „Ehrenautorenschaft“

Empfehlung 11: Autorinnen und Autoren wissenschaftlicher Veröffentlichungen tragen die Verantwortung für deren Inhalt stets gemeinsam. Eine sogenannte „Ehrenautorschaft“ ist ausgeschlossen.

(DFG, Vorschläge zur Sicherung guter wissenschaftlicher Praxis)

Überblick

TECHNISCHE UNIVERSITÄT DARMSTADT

- Plagiatur: Typen und Indikatoren
- Was leistet heutige Software?
  - Zusammenfassung des Softwaretests 2008 der FHTW Berlin
- Wie funktioniert Plagiatssoftware technisch?
  - Stand der Kunst
  - Weitere Möglichkeiten/Ausblick
- Diskussion: Was können Hochschulen tun?

Typen von Plagiatur (1)

TECHNISCHE UNIVERSITÄT DARMSTADT

(0) **Plagiatur von Autorenschaft:** Der direkte Fall, seinen eigenen Namen mit der Arbeit eines anderen zu schmücken.

(1) **Wort-für-Wort Plagiatur:** Direktes Kopieren von Phrasen oder Passagen aus einem publizierten Werk ohne Quellenangabe.

(2) **Paraphrasierende Plagiatur:** Wörter oder Satzstruktur werden abgeändert übernommen, wobei der Quelltext erkennbar bleibt.

Basierend auf Martin (1994) and Clough (2003)

Typen von Plagiatur (2)

TECHNISCHE UNIVERSITÄT DARMSTADT

(3) **Plagiatur der Form**: Die argumentative Struktur einer Quelle wird übernommen (wörtlich oder paraphrasiert).

(4) **Plagiatur von Ideen:** Originelle Gedanken einer Quelle werden ohne Abhängigkeit der Formulierung oder Argumentationsstruktur übernommen.

(5) **Plagiatur aus zweiten Quellen**: Originalquellen werden zitiert, aber durch Übernahme aus Sekundärquellen, ohne dass die Originale überprüft wurden.

Basierend auf Martin (1994) and Clough (2003)

Typische Indikatoren für Plagiatur

TECHNISCHE UNIVERSITÄT DARMSTADT

- Unerwartet fortgeschrittenes/technisches Vokabular
- Plötzliche Verbesserung des Schreibstils im Vergleich zu früheren Arbeiten

- Inkonsistenzen innerhalb eines Textes (z.B. Vokabular, Stil, Referenzen, Qualität)
- Inkohärenzen im Textfluss, die auf *cut-and-paste* hinweisen
- Fehlende Referenzen (nur im Text / nur in der Bibliographie)

- Hohe Übereinstimmung mehrerer eingereichter Texte (z.B. identische Fehler)

Basierend auf Clough (2003)

Culwins Vier-Stufen-Modell zur Erkennung von Plagiatur

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Datensammlung
   - Exemplar der zu untersuchenden Kursarbeit
   - Elektronisch, andernfalls OCR
2. Analyse (Automatisierung möglich)
   - Ähnlichkeiten zu anderen Papieren?
   - Auffälligkeiten beim Schreibstil?
   - Arbeit charakteristisch für den Studenten?
3. Bestätigung (Automatisierung schwierig)
   - Vergleich Studentenarbeit mit Originaldokument
   - Setzt Identifikation des Originaldokuments voraus
4. Untersuchung
   - Konsequenzen je nach Beweislage

Probleme

TECHNISCHE UNIVERSITÄT DARMSTADT

- Verfügbarkeit der (teils kommerziellen) Quellen
  - elektronischen Zeitschriftendatenbanken
  - hausarbeiten.de
  - Arbeiten von Mitstudierenden / aus vorigen Jahren
- Sprache (Übersetzung)
- Zeitfaktor/Aufwand

Überblick

TECHNISCHE UNIVERSITÄT DARMSTADT

- Arten von Plagiatur
- Was leistet heutige Software?
  - Zusammenfassung des Softwaretests 2008 der FHTW Berlin
- Wie funktioniert Plagiatssoftware?
  - Stand der Kunst
  - Weitere Möglichkeiten/Ausblick
- Diskussion: Was können Hochschulen tun?

Softwaretest 2008

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Prof. Dr. Debora Weber-Wulff**
  - Prof. für Media and Computing an der Fachhochschule für Technik und Wirtschaft (FHTW) Berlin
- Plagiatur-Portal http://plagiat.fhtw-berlin.de/
- Test kommerzieller und experimenteller Plagiatserkennungssoftware mit Hinblick auf Tauglichkeit im (Hoch-)Schulalltag

Softwaretest 2008 - Testmethodik

TECHNISCHE UNIVERSITÄT DARMSTADT

- **31 Testfälle** als .doc, .html (auch online), .pdf und als .txt
  - Originalaufsätze
  - Originalaufsatz, von der Autorin in die Wikipedia eingestellt
  - Übersetzungsplagiate (Babelfish, Überarbeitung, Literaturangaben)
  - Copy & Paste mit Shake & Paste
  - Halbsatzflickerei
  - Gekauft bei Hausaufgabenbörse
  - Copy & Paste eines offline Mediums (Buch, bzw. CD-ROM)
  - PDF Quellen für Copy & Paste
- **16 Systeme** (online und lokal installiert)
- **Bewertungskriterien**
  - Usability (Kostentranzparenz, Ergonomie, Einbettung in Arbeitsablauf)
  - Plagiatserkennung

Softwaretest 2008 – Bewertungskriterien (Ausschnitt)

TECHNISCHE UNIVERSITÄT DARMSTADT

Testfälle \ Punktzahl	3	2	1	0
00-schaltjahr	wenn nichts gefunden wurde	bis 10% Plagiat gemeldet	bis 25% Plagiat gemeldet oder Fußnote als Plagiat gemeldet	Große Mengen Plagiat gemeldet und/oder Warnfarbe vergeben
01-djembe	Englische Quelle gefunden	Plagiat der Seite (auf Englisch) gefunden	Plagiat der Seite (auf Deutsch) gefunden	Nichts gefunden
02-atwood	Amazon.de Quelle gefunden	Plagiatsseiten gefunden ohne Amazon	unter 20% gemeldet	Nichts gefunden
03-IETF	WZ Berlin Quelle gefunden	Quelle da, aber verwirrender Bericht		Nur wenig oder irrelevantes gefunden
08-lettau	Wikipedia gefunden	Wikipedia und Mirrors gefunden	Nur schwachsinnige Mirrors gefunden	Nichts gefunden
09-frosch	Hinweis auf schoolunity prominent, ggf. mit weitere Quellen			Nichts gefunden
10-fraktur	Wikipedia, PDF und Buch gefunden	Wikipedia und PDF gefunden	Wikipedia oder PDF oder Plagiat gefunden	Nichts gefunden

Softwaretest 2008 – Bewertungskriterien (Ausschnitt)

TECHNISCHE UNIVERSITÄT DARMSTADT

Testfälle \ Punktzahl	3	2	1	0
00-schaltjahr	wenn nichts gefunden wurde	bis 10% Plagiat gemeldet	bis 25% Plagiat gemeldet oder Fußnote als Plagiat gemeldet	Große Mengen Plagiat gemeldet und/oder Warnfarbe vergeben
01-djembe	Englische Quelle gefunden	Plagiat der Seite (auf Englisch) gefunden	Plagiat der Seite (auf Deutsch) gefunden	Nichts gefunden
02-atwood	Amazon.de Quelle gefunden	Plagiatsseiten gefunden ohne Amazon	unter 20% gemeldet	Nichts gefunden
03-IETF	WZ Berlin Quelle gefunden	Quelle da, aber verwirrender Bericht		Nur wenig oder irrelevantes gefunden
08-lettau	Wikipedia gefunden	Wikipedia und Mirrors gefunden	Nur schwachsinnige Mirrors gefunden	Nichts gefunden
09-frosch	Hinweis auf schoolunity prominent, ggf. mit weitere Quellen			Nichts gefunden
10-fraktur	Wikipedia, PDF und Buch gefunden	Wikipedia und PDF gefunden	Wikipedia oder PDF oder Plagiat gefunden	Nichts gefunden

Softwaretest 2008 – „Sieger“

Platz 1 Indigo Stream Technologies Ltd., **Copyscape Premium**.

- Verwendet Google-API
- 3 min Laufzeit pro Text
- Hash-Verfahren, kein Stringvergleich
  - Fontwechsel hat Auswirkungen
- Erreichte im Test 70 von 80 Punkten
- − Es kann nur eine URL oder Datei pro Zeiteinheit geprüft werden
- − keine Quantifizierung des Übereinstimmungsgrads
- \+ jeder Test kostet 5 US-Cent

Softwaretest 2008 – „Sieger“

TECHNISCHE UNIVERSITÄT DARMSTADT

Platz 2 Plagiarism-Detector

- Verwendet nur online-Quellen
- Erreichte im Test 68 Punkte
  - Dubiose Supportadresse
  - Installation kompliziert (Installiert ein Trojaner!)
  - pdf-Support funktioniert nicht
- Kosten zwischen 50 USD und 100 USD je nach Umfang (mit/ohne ppt-Support etc.)

Softwaretest 2008 – Alle Systeme im Überblick

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Sehr gute Systeme**
- **Gute Systeme**
  1. Copyscape Premium
  2. Plagiarism-Detector
  3. Copyscape Free
  4. Urkund
  5. Docoloc - PlagAware (neu)
  6. Ephorus
- **Befriedigende Systeme**
  7. SafeAssign
  8. Strikeplagiarism
  9. PlagiatCheck
  10. AntiPlag
  11. PlagAware (alt)
- **Ausreichende Systeme**
  12. Turnitin
  13. XXXX
- **Nicht-ausreichende Systeme**
  15. Plagiarism-Finder
  16. Turnitin Global
- **Abgebrochene Tests**
  - ArticleChecker
  - CatchItFirst
  - checkforplagiarism.net
  - paperseek
  - WebMasterLabor

Softwaretest - Diskussion

TECHNISCHE UNIVERSITÄT DARMSTADT

- Mehrzahl der Systeme durchsucht online-Inhalte
  - Starke Verbesserung im Vergleich zu früheren Tests
  - Verwendung etwas bequemer als Suche „per Hand“
  - Qualität/Ergonomie der Software noch nicht optimal

- Es werden nur wenige Arten von Plagiatur erkannt (Abschreiben, einfaches Editieren)
  - Gekaufte Quelle war über Google zu finden
  - Übersetzte Quellen wurden nicht erkannt
  - i.d.R. keine Kollusionserkennung

Softwaretest 2008 – Kollusionserkennung

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Originalwerk und vier abgeschriebene Versionen**
  1. jeweils der erste und letzte Satz verändert
  2. ersten Absatz stark verändert
  3. über den ganzen Text einzelne Wörter durch Synonyme ersetzt
  4. andere Schriftart in Word verwendet, der Text blieb jedoch identisch
- **Gute Systeme**
  - **JPlag** (Universität Karlsruhe)
    - eigentlich für Programmcode-Vergleich entwickelt
    - Schnell (42 txt < 1 Minute) und kostenlos
  - **Turnitin**
    - einziges Plagiatstest-System, das Kollusionen zu erkennt, wenn die Aufsätze in der Datenbank gespeichert werden.
  - **WCopyFind**
    - Viele Einstellungsmöglichkeiten, schnell

Softwaretest - Diskussion

TECHNISCHE UNIVERSITÄT DARMSTADT

- Datenbank nötig, wenn Kollusion und Plagiatur aus nicht-online-Quellen geprüft werden sollen
  - Datenschutz-Fragen
  - Gefahr, dass die geprüften Texte gespeichert werden

- Verhältnismäßigkeit:
  - Eingriff nur mit Begründung; Art. 2 GG (Allg. Persönlichkeitsrecht);
  - Die systematische Überprüfung aller Arbeiten kann als Generalverdacht gegen alle Studierenden aufgefasst werden.
  - Haushaltsrecht: Verhältnis Aufwand/Nutzen;

- Geeignetheit:
  - Software-Angebote zur Verschleierung von Plagiaten;
  - Bei Identität von Arbeiten Urheber nicht feststellbar.

(Behrendt 2007)

Überblick

TECHNISCHE UNIVERSITÄT DARMSTADT

- Arten von Plagiatur
- Was leistet heutige Software?
  - Zusammenfassung des Softwaretests 2008 der FHTW Berlin
- Wie funktioniert Plagiatssoftware?
  - Stand der Kunst
  - Weitere Möglichkeiten/Ausblick
- Diskussion: Was können Hochschulen tun?

Verbreiteter Ansatz: Stringvergleich

TECHNISCHE UNIVERSITÄT DARMSTADT

- Wortfenster der Länge n (N-Grams) werden verglichen
- Länge der Wortfenster wird empirisch festgelegt

  a. *Am Sonntag hat die große Koalition sich auf eine Mehrwertsteuererhöhung geeinigt*
  b. *Die große Koalition aus SPD und CDU hat sich am Sonntag auf eine Mehrwertsteuererhöhung geeinigt*

- Je größer das n, um so geringer die Wahrscheinlichkeit, dass zwei Autoren unabhängig dasselbe N-Gramm verwenden
- Problem: geeignete Schwellenwerte finden

Eindeutigkeit von N-Grammen
(aus Clough 2003)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Zahlen aus 769 Texten im METER Textkorpus

N (words)	N-gram occurrences (tokens)	Distinct n-grams (types)	% distinct n-grams	% distinct n-grams in 1 file
1	137204	14407	11	39
2	248819	99682	40	67
3	248819	180674	73	82
4	257312	214119	85	90
5	251429	226369	90	93
6	250956	231800	92	94
7	250306	234600	94	95
8	249584	236310	95	96
9	248841	237409	95	97
10	289610	278903	96	97

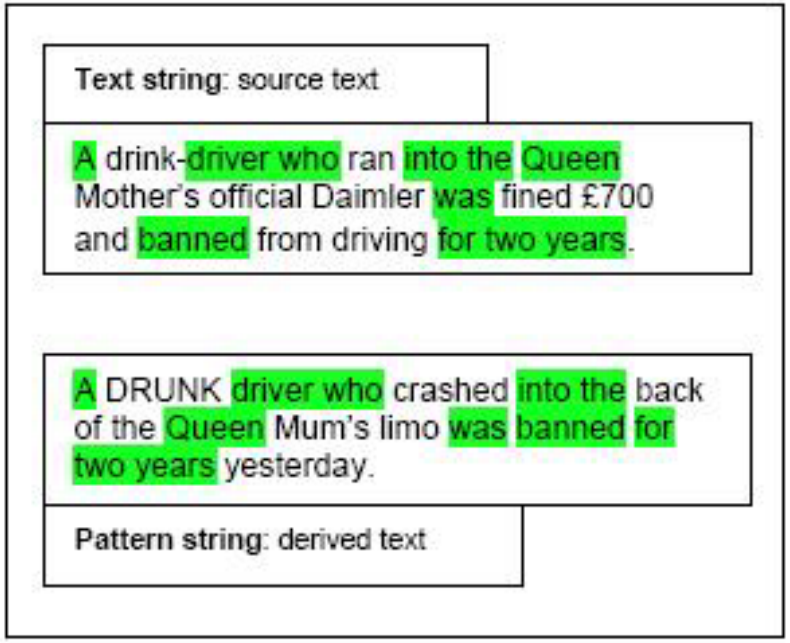
Table 1 Uniqueness of consecutive n-word sequences (n-grams) as *n* increases from 1-10 words

Anderer Ansatz: Berechnung der längsten gemeinsamen Zeichenkette

TECHNISCHE UNIVERSITÄT DARMSTADT

- Greedy String Tiling (Wise,1993) ist ein Algorithmus, der die **maximale Abbildung Teile zweier Texte** so berechnet, dass sich die Wortsequenzen **nicht überlappen**

- Vorteil: Man muss keine Länge n apriori festlegen



Weiterverarbeitung der GST-Ergebnisse

TECHNISCHE UNIVERSITÄT DARMSTADT

- Der Output von GST im Beispiel ist die Liste
  [for two years], [driver who], [into the], [a], [queen], [was] und [banned].

- Verschiedene Maße können nun angewendet werden, um Plagiatur zu modellieren
  - Mindest- und Maximallänge der Sequenzen
  - Durchschnittliche Länge
  - Die Verteilung der Längen

- Ziel: Ein Ähnlichkeitsmaß für Plagiatur zu entwickeln

- Eine Herausforderung: Erkennen, was Original und was Plagiat ist.

Beispiel-Aufteilung von abgeleitetem und Originaltext

TECHNISCHE UNIVERSITÄT DARMSTADT

- Empirische Beobachtung:
  - Abgeleitete Texte (oben) enthalten längere matchende Teilstrings
  - Die Aufteilung von Original und Ableitung unterscheiden sich zumeist deutlich

Maschinelles Lernen zur Plagiatserkennung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Eingabe: Dokumente und ihre Merkmale
  - z.B. Dokumentlänge, Anzahl und durchschnittliche Länge der Sequenzen
- Ziel: Computermodell soll Original und Plagiat unterscheiden
  - ***Supervised learning:*** Computer wird auf bereits ausgezeichneten Dokumenten (Orig./Plag) “trainiert”
    - Nachteil: große Datenmengen nötig (1000ende von Beispielen)
  - ***Unsupervised learning***: Computer soll “ähnliche Cluster” finden
    - Konkrete Anweisung: Teile die Dokumente in zwei Klassen
    - Hoffnung: Die eine enthält nur Originaldokumente und die andere Plagiate
- Überprüfung: Stichproben

Ausblick/Weitere Möglichkeiten

TECHNISCHE UNIVERSITÄT DARMSTADT

Ziel: Verfahren intelligenter machen, also robuster gegenüber “Edits”, die beim Matching nicht erkannt werden

- Erlaube kleine Lücken (Löschen einzelner Wörter)
- Erlaube das Einsetzen von Funktionswörtern und “Füllwörtern”
- Erkenne Wortersetzungen (Wörterbücher, Thesauri)
- Erkenne das Einsetzen von Fachvokabular
- Erkenne Variation in der Wortstellung (gerade für D. interessant)

Ausblick/Weitere Möglichkeiten

TECHNISCHE UNIVERSITÄT DARMSTADT

- Autorenerkennung ist in der theoretischen Linguistik (auch: Forensik) etabliert
  - Verfahren könnten auf den Computer übertragen werden
    - Bsp.: Ist der Text in sich homogen (Satzlänge, Vokabular, Stilmittel)?

- Allgemeine Probleme
  - Texte sind sehr unterschiedlich strukturiert (Essay / Technischer Bericht)
  - Manche Inhalte sind nicht-textueller Natur (Formeln, Referenzen, Tabellen, Grafiken)

- Verwandte sprachtechnologische Einsatzmöglichkeiten
  - Automatisches Assessment (Lernstandskontrolle)
  - Informationszugriff (Vorschläge, Zusammenfassungen, Verlinkung)
  - …

Überblick

TECHNISCHE UNIVERSITÄT DARMSTADT

- Arten von Plagiatur
- Was leistet heutige Software?
  - Zusammenfassung des Softwaretests 2008 der FHTW Berlin
- Wie funktioniert Plagiatssoftware?
  - Stand der Kunst
  - Weitere Möglichkeiten/Ausblick
- Diskussion: Was können Hochschulen tun?

Diskussion: Was können Hochschulen tun?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Vorbild sein
- Ab Semester 1 wissenschaftliche Qualitätsstandards vermitteln
- Wachsam sein
  - Software kann helfen
    - Für Online-Quellen: Suchmaschinen
    - Innerhalb von Kursen: Rechtslage (Datenspeicherung) beachten
- Handhabe bei Plagiatur klären und Vermitteln
- Open Content?

Referenzen

- **Lancaster, T. and Culwin, F. (2001):** Towards an error free plagarism detection process. In *Proceedings of the 6th Annual Conference on innovation and Technology in Computer Science Education* (Canterbury, United Kingdom). ITiCSE '01. ACM, New York, NY.
- **Berendt, Bettina (2007):** Anti-Schummel-Software oder Hilfe bei der wissenschaftlichen Ausbildung? Plagiatsdetektion und -prävention. *CMS Journal, 29, Sonderheft: Facetten von Bologna*.
- **Clough, Paul (2003):** Old and new challenges in automatic plagiarism detection, National UK Plagiarism Advisory Service.
- **Martin, Brian (1994):** Plagiarism: a misplaced emphasis. In *Journal of Information Ethics*, 3:2(36-47)
- **Sattler, Sebastian (2007):** Plagiate in Hausarbeiten Erklärungsmodelle mit Hilfe der Rational Choice Theorie, SOCIALIA - Studienreihe soziologische Forschungsergebnisse, Hamburg.
- **Wise, Michael (1993):** String Similarity for Greedy String Tiling and Running Karp-Rabin Matching. *Technical report available at* ftp://ftp.cs.su.oz.au/ michaelw/doc/RKR_GST.ps. Department of Computer Science, University of Sydney.