Informationstechnische Aspekte des *Historical Text Re-use* - Zusammenfassung -

Der Fakultät für Mathematik und Informatik
der Universität Leipzig
eingereichte

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR rerum naturalium
(Dr. rer. nat.)

im Fachgebiet

Informatik

vorgelegt

von Dipl.-Inf. Marco Büchler
geboren am 24. Juli 1978 in Eilenburg

Leipzig, den 4. Februar 2013

Gegenstand der Arbeit

Was ist *Text Re-use*? *Text Re-use* beschreibt die mit unterschiedlichen Absichten mündliche und schriftliche Wiedergabe von Textinhalten. Diese können im Sinne einer Definition das Anerkennen einer Autorität aber auch das Wiedergeben einer besonders interessanten Information sein. Während der Fokus dieser Arbeit auf dem Erstellen eines *Hypertextes* durch eine *Text Re-use Analysis* liegt, sind die *PageRanking*-Technik oder auch bibliometrische Analysen weiterführende Anwendungen. Im Kontext derartiger Einsatzmöglichkeiten kann auf historischen Dokumenten, die dieser Arbeit zugrunde liegen, durch eine automatische Analyse eine noch nie zuvor erstellte Breite von Zitierabhängigkeiten erstellt werden, welche heutzutage Aufschluss darüber geben, was in früheren Zeiten als wichtig erachtet worden ist, auch wenn es in der Gegenwart für Sprachen, wie dem Altgriechischen oder dem Latein, keine Muttersprachler mehr gibt.

Stand der Forschung

In der Plagiarismuserkennung, einer modernen Anwendung von *Text Re-use*, werden meist einfache *Ngramm*-Ansätze eingesetzt. Diese Form einer Abtastung eines Textes bietet in erster Linie den Vorteil, dass die benötigte Rechenzeit relativ klein bleibt. Ferner genügt dieser Ansatz, um ein einfaches *Copy & Paste* zu erkennen.

Außerhalb des Plagiarismus stellt sich der Forschungsstand so dar, dass nahezu beliebig Daten und Algorithmen kombiniert werden. Die Ergebnisse geben datenspezifische Charakteristika wieder und sind somit oft nicht auf andere Daten reproduzierbar. Der Forschungsstand reflektiert somit mehr Insellösungen als eine ganzheitliche Sicht auf das Thema.

Ganzheitliche Sicht auf *Text Re-use*

In Kapitel 2 wird die derzeit vollständigste Systematisierung des *Text Re-use* vorgenommen. Dies umfasst zwei wesentliche Aspekte:

- Es werden insgesamt 45 verschiedene Typisierungen von Textstellen, nachfolgend auch *Meme* im Sinne eines Gedanken oder Gedankensplitters genannt, eingeführt, welche in der Regel wiederverwendet werden. Entsprechende typisierte *Meme* reichen nur beispielhaft von *Sprichwort*, über *Schlachtruf* und *Vers* bis hin zur *Legende*.
- Es wird eine Systematik zu verschiedenen *Re-use Styles* definiert, welche beschreibt, wie ein entsprechendes *Meme* wiederverwendet wird. Das kann zum Beispiel ein wortwörtliches Zitat aber auch eine Paraphrase oder Allusion sein.

Das Ziel dieser ganzheitlichen Sicht besteht darin, grundlegende Eigenschaften der *Meme* sowie der *Re-use Styles* zu definieren. Während ein *Meme*, wie z. B. eine *Redewendung*, eher kurz und syntaktisch fest verwendet wird, ist es beim größeren *Meme Legende* üblich, dieses mündlich und damit wesentlich freier wiederzugeben.

Während die Typisierung der verschiedenen *Meme* die Frage aufwirft, warum bestimmte Textinhalte wiederverwendet werden, gibt die zweite Systematik des *Re-use Styles* Aufschluss darüber, wie jeder persönlich andere Inhalte wiedergibt.

Sowohl die Typisierung der verschiedenen *Meme* mit ihren unterschiedlichen Charakteristika als auch die Systematik der *Re-use Styles* reflektieren eine *Data Diversity*, welche eine Herausforderung sowohl für die *Text Re-use Analysis* aber auch für deren *Evaluation* aus ganzheitlicher Sicht bedeutet, da es keinen *Gold Standard* gibt, welcher sowohl alle möglichen *Meme* als auch die verschiedenen *Re-use Styles* adäquat repräsentiert.

Forschungsfragen

Aus ganzheitlicher Sicht ergeben sich somit für diese Arbeit die folgenden Forschungsfragen:

- Im Kontext der verschiedenen *Re-use Styles* muss die Frage danach gestellt werden, bis zu welchem Grad der Veränderung ein *Text Re-use* automatisch noch erkannt werden kann.
- Wie kann eine *Text Re-use Analysis* so gestaltet werden, dass sie auch für unterschiedliche *Meme* mit verschiedenen Charakteristika gleich gut funktioniert?
- Wie können Veränderungen eines wiederverwendenden Autors systematisch bestimmt und extrahiert werden?
- Wie kann das Ergebnis einer *Text Re-use Analysis* in einer *Digital Library*[1] in Anbetracht der *Data Diversity* ganzheitlich evaluiert werden?

Untersuchte *Digital Libraries*

Im Gegensatz zum *Text Re-use* auf modernen Texten, wie in der Plagiarismusforschung, wird eine Analyse auf historischen Dokumenten dadurch erschwert, dass keine vereinheitlichte Schreibweise angenommen werden kann, da neben autorspezifischen Aspekten auch Sprachevolution, Dialekte, semantischer Wandel von Konzepten aber auch verschiedene Varianten, verursacht durch eine weniger reglementierte und in den Anfängen der Verschriftlichung von Texten nicht existenten Rechtschreibung, den Schreibprozess begleiten.

Diese Arbeit vereint somit im Rahmen des *Historical Text Re-use* sowohl die *Data Diversity*, bedingt durch verschiedene *Meme* und *Re-use Styles*, mit der großen sprachlichen Vielfalt von historischen Dokumenten. Als Datenbasis liegen dieser Arbeit drei verschiedene *Digital Libraries* zugrunde:

- *Perseus Digital Library*: Diese Datenbasis wird eingesetzt, um sehr kurzen *Text Reuse* auf altgriechischen Werken mit einem hohen Maß an sprachlicher Vielfalt zu analysieren.
- *Bibelversionen*: Es werden insgesamt sieben verschiedene englischsprachige Bibelversionen verwendet. Ausgehend von der *King James Version* aus dem 16. Jh., welche noch sehr alte und archaische Wortformen enthält, werden die Verse auch mit anderen Versionen der Bibel verglichen, welche bspw. der hebräischen Satzsyntax folgen oder den Inhalt eines Verses mit möglichst einfacher Sprache wiedergeben.
- *Sammlungen von Redewendungen*: Weiterhin erfolgt die Analyse zweier kleiner deutschsprachiger Sammlungen von Redewendungen, welche einen Ursprung im Mittelalter bzw. einen Bezug zur Bibel haben.

Untersuchungsmethodik und Lösungsansatz

Da die *Data Diversity* aus informationstechnischer Sicht nicht mit einem einzelnen Algorithmus bzw. einer kleinen Menge von Ansätzen abgedeckt werden kann, wird in Kapitel 3 die *7-Level-Architektur* des *Historical Text Re-use* vorgestellt. Diese Architektur kann als ein modulares Konzept verstanden werden, um die *Text Re-use Analysis* auf die verschiedenen

[1] Als *Digital Library* wird eine digitale Kollektion von Texten verstanden.

Bedürfnisse, bedingt durch spezielle Eigenschaften von *Meme*, unterschiedlichen *Re-use Styles* aber auch verschiedenen Sprachvarianten, entsprechend anzupassen. Die einzelnen Level entsprechen den sieben Unteraufgaben *Segmentation*, *Preprocessing*, *Featuring*, *Selection*, *Linking*, *Scoring* und *Postprocessing*. In Kapitel 3 werden zu jedem Level in einem separaten Abschnitt entsprechende Implementierungen sowohl ausführlich vorgestellt als auch systematisiert. Zur Abgabe dieser Dissertation stehen in der *TRACER*-Implementierung, welche die *7-Level-Architektur* umsetzt, insgesamt über eine Million Kombinationsmöglichkeiten der verschiedenen Ansätze der einzelnen Level zur Verfügung.

Sowohl die drei genannten Forschungsfragen als auch die aufgezeigte *Data Diversity* des *Historical Text Re-use* werden im Rahmen der Dissertation als hinreichende Motivation verstanden, den *Historical Text Re-use* in Shannon's *Noisy Channel Theorem* einzubetten. In diesem Kontext kann ein Original- bzw. zitierter Autor als *Source* und ein wiederverwendender Autor als *Target* verstanden werden. Der *Noisy Channel* stellt ein unbekanntes Modell von Modifikationen, den äußeren Einflüssen, dar.

In Kapitel 4 wird das *Noisy Channel Model* dazu eingesetzt, ein zufälliges und rein künstliches Störsignal zum *Noisy Channel* hinzuzufügen, so dass eine *Randomised Digital Library* entsteht. Es werden insgesamt fünf Klassen von Randomisierungstechniken, die künstlichen Störsignale, im Sinne eines Turingtests vorgestellt, welche unterschiedliche Schwierigkeitsgrade einer rein quantitativen Evaluierung mit sich bringen. Für diese quantitative Evaluierung, die *Noisy Channel Evaluation*, wird der neuartige *Score* der *Mining Ability* eingeführt. Die *Mining Ability* setzt hierbei das Ergebnis einer *Text Re-use Analysis* auf einer *Digital Library* mit dem Resultat einer durch ein künstliches Störsignal veränderten *Randomised Digital Library* ins Verhältnis, wodurch nicht nur Parameter optimiert sondern auch verschiedene Sprachmodelle vollautomatisch und bzgl. des Ergebnisses ganzheitlich sowie ohne *Gold Standard* evaluiert werden können.

In Kapitel 5 wird der *Noisy Channel* als Modell eingesetzt, um historisch paradigmatische Relationen systematisch zu bestimmen. Das ist insbesondere unter Berücksichtigung der großen Zeitfenster von geisteswissenschaftlichen Texten von Interesse, da sich semantische Beziehungen von Konzepten im Laufe der Zeit verändert haben.

Ergebnisse

Die Ergebnisse dieser Arbeit sind sehr vielschichtig und umfassen neben Ergebnissen von Evaluierungen, auch Erfahrungen innerhalb der *eHumanities* sowie der entsprechenden Grundlagenarbeit. Im Detail können die Ergebnisse wie folgt zusammengefasst werden:

Es wird im einführenden Kapitel der Dissertation das Paradigma *ACID for the eHumanities* vorgestellt. *ACID* ist hierbei eine Abkürzung für *Acceptance*, *Complexity*, *Interoperability* und *Diversity*. Diese vier Säulen werden als Aspekte vorgestellt, denen sich die Informatik in der Zusammenarbeit mit den Geisteswissenschaften stellen muss. Der Fokus der Arbeit liegt auf der *Diversity* aber auch Aspekte der *Acceptance* und *Complexity* werden ausführlich verdeutlicht.

In Kapitel 4 wird neben der Einführung der *Noisy Channel Evaluation* auch aufgezeigt, welche statistischen Probleme probabilistische Sprachmodelle begleiten. Während probabilistische Sprachmodelle das *Gesetz der großen Zahlen* und somit eine hinreichend große Auftretenswahrscheinlichkeit voraussetzen, folgen verschiedene Charakteristika natürlicher Sprache einem *Power Law*, wie dem Zipfschen Gesetz, so dass für den *Long Tail* dieser Verteilung eine geringe Frequenz zugrunde liegt, woraus letztlich ein statistisches Problem resultiert. Im Detail kann so gezeigt werden, dass der eingeführte *Score* der *Mining Ability* bei zunehmender Größe einer *Digital Library* nach Erreichen eines Maximums wieder sinkt. Das resultiert daraus, dass mit zunehmender Größe der *Digital Library* vermehrt

aus Rauschen als Neuem "gelernt" wird. Auch wenn Kapitel 4 das auf den *Text Re-use* einschränkt, so sind die Ergebnisse einfach auf andere probabilistische Sprachmodelle adaptierbar. Insbesondere wird der Widerspruch des *Gesetzes der großen Zahlen*, welches den auf Wahrscheinlichkeiten aufsetzenden Sprachmodellen implizit zugrunde liegt, und den oftmals sehr seltenen Ereignissen beim Umgang mit natürlichsprachlichen Texten deutlich.

In Kapitel 5 wird weiterhin gezeigt, dass es kein *Text Re-use Model* gibt, welches in jedem Szenario optimale Ergebnisse liefert. Basierend auf sieben Bibelversionen mit unterschiedlichen Bezügen untereinander, wird verdeutlicht, dass sich nicht nur die Algorithmen der *7-Level-Architektur* unterscheiden können, sondern auch entsprechende Schwellwerte.

Im Rahmen der Arbeit werden zwei rein quantitative Evaluierungsgrößen, die *Text Re-use Compression* sowie die *Noisy Channel Evaluation*, eingeführt. In Kapitel 5 wird gezeigt, dass es eine signifikante Korrelation zu existierenden Evaluierungsgrößen gibt, welche jedoch einen *Gold Standard* oder zumindest eine Evaluierungsgrundlage benötigen. Einerseits gibt es eine nach Pearson sehr starke Korrelation zwischen dem *Recall* und der *Text Re-use Compression*. Andererseits wird auch gezeigt, dass das *F-Measure* sowie die im Rahmen dieser Arbeit eingeführte *Noisy Channel Evaluation* sehr vergleichbare Evaluierungsergebnisse erzeugen. Das wird im Rahmen einer *System Evaluation* in Kapitel 5 anhand der sieben Bibelversionen in insgesamt 504 verschiedenen Experimenten dargestellt.

Beitrag zur Forschung

Neben den aufgezeigten Ergebnissen stellt diese Arbeit Grundlagenforschung sowohl in der Systematisierung des *Text Re-use* aber auch bei der Evaluierung von Ergebnissen dar. Wie eingangs zum Forschungsstand umrissen wurde, verlieren sich derzeit viele Arbeiten in der nahezu beliebigen Kombination aus Daten und Algorithmen. Mit dieser Arbeit wird ein Evaluierungsszenario vorgestellt, welches es ermöglicht, auch ohne *Gold Standard* das Ergebnis zu bewerten. Somit wird das Resultat nicht mehr durch unterschiedliche Überlappungsgrade zwischen *Digital Library* und *Gold Standard* verfälscht.

Des Weiteren geht mit dieser Arbeit ein Paradigmenwechsel einher. Während in der Automatischen Sprachverarbeitung *Text Re-use* bisher aus einer "*1-Algorithmus-Sicht*" betrachtet wird, zeigen die Ergebnisse aus Kapitel 5 auf, dass zukünftig stärker der paarweise Vergleich zweier Werke im Forschungsvordergrund stehen sollte. Das geht damit einher, dass jeder Mensch einen eigenen *Re-use Style* besitzt, so dass durch das paarweise Vergleichen die menschlichen Individualitäten im Fokus der *Text Re-use Analysis* stehen. Deshalb wird vorgeschlagen, die Einzelergebnisse der werkweisen Vergleiche anschließend zu einem *Hybrid Text Re-use Graph* zusammenzusetzen. Mit der *Noisy Channel Evaluation* sowie der *Text Re-use Compression* stehen nun weiterführend auch vollautomatische Evaluierungstechniken zur Verfügung, so dass eine wesentlich präzisere *Text Re-use Analysis* möglich ist.

Perspektive

Entgegen modernen Anwendungen des *Text Re-use*, wie dem Plagiarismus, kann der *Historical Text Re-use* als ein nützliches Instrument verstanden werden, welches nicht nur Evidenzen von Transferwegen, sondern vielmehr auch einen fundamentalen Teil des sprachlich-kulturellen Erbes der Menschheit darstellt. Aus der Vielfalt des *Historical Text Re-use* ergeben sich für die Informatik im Rahmen der *eHumanities* vielschichtige Herausforderungen, die Gegenstand dieser Arbeit sind. Im Detail bedeutet das einen Paradigmenwechsel vom Pragmatismus im Vergleich von Sprachmodellen hin zur bestmöglichen Vollständigkeit.