Delta vs. N-Gram-Tracing: Wie robust ist die Autorschaftsattribuierung?

Proisl, Thomas
thomas.proisl@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutschland

Evert, Stefan
stefan.evert@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutschland

Die Autorschaftsattribuierung, also die Zuweisung von Texten unbekannter oder umstrittener Autorschaft zu ihrem wahren Autor, hat vielfältige Anwendungen beispielsweise in der Literatur- und Geschichtswissenschaft oder der forensischen Sprachwissenschaft. Eine populäre Methode zur Autorschaftsattribuierung ist die Anwendung von Deltamaßen (Burrows 2002; Argamon 2008) wie zum Beispiel Cosine-Delta (Smith und Aldridge 2011). Deltamaße verwenden die n häufigsten Wörter im Korpus, standardisieren die Frequenzen auf z-Werte und wenden ein Abstandsmaß, im Fall von Cosine-Delta den Kosinusabstand, an. Typischerweise schließt sich die Anwendung eines (hierarchischen) Clusterverfahrens an, das Texte des selben Autors zusammengruppiert.

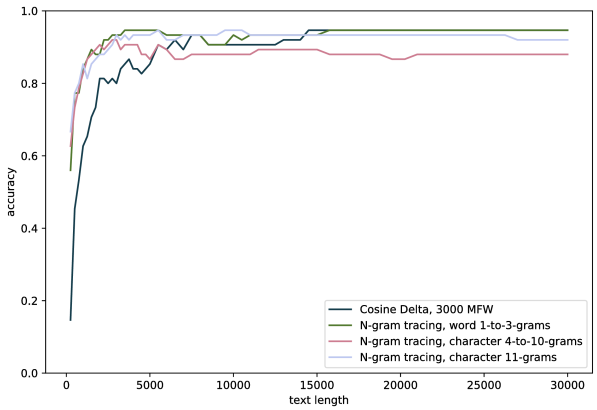
Eine neue Methode zur Autorschafts-attribuierung ist das sogenannte N-Gram-Tracing (Grieve et al., in Begutachtung). Hierbei werden aus dem zu klassifizierenden Text alle Wort- oder Buchstaben-N-Gramme einer bestimmten Länge extrahiert. Der Text wird dann dem Autor zugewiesen, der im Vergleichskorpus die meisten dieser N-Grammtypen verwendet. Die Häufigkeit der N-Gramme spielt dabei keine Rolle, es geht nur darum, wie viele N-Gramme aus dem zu klassifizierenden Text auch im Vergleichskorpus auftauchen.

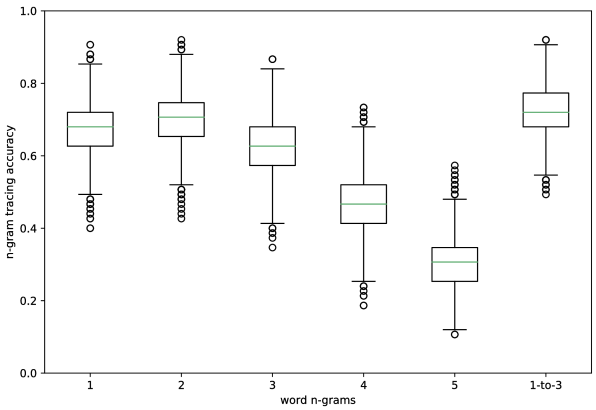
Wenn Methoden zur Autorschaftsattribuierung angewandt werden sollen um tatsächlich eine strittige Autorschaftsfrage zu klären, ist es sehr wichtig die Zuverlässigkeit und Robustheit der Verfahren abschätzen zu können, schließlich gibt es eine ganze Reihe von Einflussfaktoren. Kritisch sind zum Beispiel die folgenden Fragen: Welchen Einfluss haben die Länge des zu klassifizierenden Textes und die Größe des Vergleichskorpus auf die Genauigkeit der Autorschaftsattribuierung? Gibt es für die beiden Verfahren eine Mindesttextlänge, die nicht unterschritten werden sollte? Wie stark werden die Verfahren durch autor- und werkspezifische Eigenheiten beeinflusst? Ist die Genauigkeit der Autorschaftsattribuierung robust in Bezug auf die Zusammensetzung des Vergleichskorpus oder kann die Auswahl der Autoren und Texte das Ergebnis beeinträchtigen?

Um diese Fragen zumindest teilweise beantworten zu können, führen wir eine Reihe von Evaluationsexperimenten durch. Um die Ergebnisse des N-Gram-Tracings besser mit denen von Delta vergleichen zu können, führen wir auf den Deltaabständen zwischen den Texten kein Clustering sondern eine nearest-neighbor-Klassifikation durch, d.h. wir weisen den zu klassifizierenden Text dem Autor des Textes mit dem geringsten Abstand zu. Im Einzelnen handelt es sich um zwei Kürzungs- und zwei Samplingexperimente. Datengrundlage für die Kürzungsexperimente sind die deutschen, englischen und französischen Romankorpora, die unter anderem von Jannidis et al. (2015) und Evert et al. (2017) verwendet wurden. Jedes Korpus besteht aus je drei Romanen von 25 Autoren, also aus 75 Romanen. Für das erste Kürzungsexperiment wird die Größe des Vergleichskorpus stabil gehalten und nur der zu klassifizierende Text gekürzt. Für Delta wird zusätzlich die Anzahl der verwendeten häufigsten Wörter variiert. Im zweiten Kürzungsexperiment werden sowohl der zu klassifizierende Text als auch das Vergleichskorpus gekürzt. Über ein leave-one-out-Verfahren werden alle Texte im Korpus klassifiziert um die Genauigkeit der Verfahren zu ermitteln.

Für die Samplingexperimente verwenden wir eine Sammlung von 1018 deutschen Romanen aus dem langen 19. Jahrhundert. Alle Texte wurden von Muttersprachlern verfasst. Für das erste Samplingexperiment ziehen wir 5000 zufällige Stichproben von 25 Autoren und je drei Romanen (die Zusammensetzung der einzelnen Stichproben ist also vergleichbar mit den oben erwähnten Romankorpora). Für das zweite Samplingexperiment beschränken wir uns auf die 25 Autoren, die in unserer Sammlung mit den meisten Romanen vertreten sind, und ziehen 5000 zufällige Stichproben von je drei Romanen pro Autor (also ebenfalls 25×3 Texte). Für jede Stichprobe ermitteln wir über ein leave-one-out-Verfahren die Genauigkeit der beiden Verfahren.

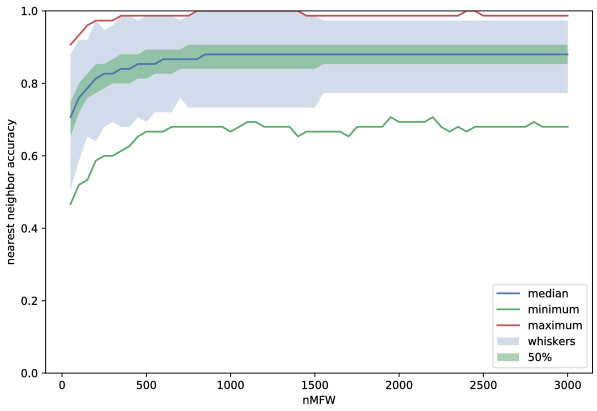
Aus Platzgründen berichten wir an dieser Stelle nur knapp die Ergebnisse des ersten Kürzungsexperiments und des ersten Samplingexperiments und beschränken und dabei auf auf die deutschen Daten. Die Ergebnisse des ersten Kürzungsexperiments, in dem nur der zu klassifizierende Text gekürzt wird, sind in der folgenden Abbildung dargestellt:





Wir vergleichen Cosine-Delta auf Basis der 3000 häufigsten Wörter mit N-Gram-Tracing auf Basis von Wort-1-bis-3-Grammen, Zeichen-4-bis-10-Grammen und Zeichen-11-Grammen. Bis zu einer Textlänge von 5000 Tokens liefern alle N-Gram-Tracing-Varianten bessere Ergebnisse als Delta, für längere Texte funktionieren Delta und N-Gram-Tracing auf Wort-N-Grammen am besten. Bei weniger als 2000 Tokens brechen die Ergebnisse für Delta ein, bei weniger als 1000 Tokens auch die für N-Gram-Tracing.

Die Ergebnisse des ersten Samplingexperiments zeigen, dass die Klassifikationsgenauigkeit bei beiden Verfahren großen Schwankungen unterworfen ist. Hier die Ergebnisse für Cosine-Delta:



Die Grafik zeigt, dass ab ca. den 1000 häufigsten Wörtern zwar im Mittel eine Klassifikationsgenauigkeit von rund 85% erreicht wird, allerdings mit enormen Schwankungen zwischen knapp über 60% und knapp unter 100%.

Die Ergebnisse für N-Gram-Tracing auf Basis von Wort-N-Grammen sehen ähnlich aus:

Durch die Kombination von Wort-1- bis Wort-3-Grammen wird zwar eine mittlere Klassifikationsgenauigkeit von über 70% erreicht, aber auch hier mit enormen Schwankungen.

Die Ergebnisse zeigen, dass N-Gram-Tracing auf kurzen Texten besser funktioniert als Cosine-Delta, allerdings werden für beide Verfahren längere Texte benötigt, als häufig verwendet werden. Die Wahl der Autoren im Vergleichskorpus und auch, wie das Poster zeigen wird, die Wahl der einzelnen Werke haben einen enormen und schwer vorhersehbaren Einfluss auf die Qualität der Autorschaftszuschreibung, deren Genauigkeit ohne weiteres um 20 Prozentpunkte schwanken kann. Im Licht dieser Erkenntnisse ist es durchaus fraglich, wie valide und generalisierbar bisherige Forschungsergebnisse auf dem Gebiet der Autorschaftsattribuierung sind.

Bibliographie

Argamon, Shlomo (2008): „Interpreting Burrows' delta: Geometric and probabilistic foundations“. In: *Literary and Linguistic Computing* 23/2: 131–47. https://doi.org/10.1093/llc/fqn003

Burrows, John (2002): „'Delta'—A measure of stylistic difference and a guide to likely authorship“. In: *Literary and Linguistic Computing* 17/3: 267–87. https://doi.org/10.1093/llc/17.3.267.

Evert, Stefan / Proisl, Thomas / Jannidis, Fotis / Reger, Isabella / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2017): „Understanding and explaining Delta measures for authorship attribution.“ In: *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx023.

Grieve, Jack / Carmody, Emily / Clarke, Isobelle / Gideon, Hannah / Heini, Annina / Nini, Andrea / Waibel, Emily (in Begutachtung): „Attributing the Bixby Letter using n-gram tracing“. Eingereicht bei *Digital Scholarship in the Humanities* am 26. Mai 2017.

Jannidis, Fotis / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2015): „Improving Burrows' Delta – An empirical evaluation of text distance measures“. In:

Digital Humanities 2015: Conference Abstracts. http://dh2015.org/abstracts.

Smith, Peter W. H. / Aldridge, W. (2011): „Improving authorship attribution: Optimizing Burrows' delta method". In: *Journal of Quantitative Linguistics* 18/1: 63–88. https://doi.org/10.1080/09296174.2011.533591.