

Sentimentanalyse in unstrukturierten Texten (am Bsp. literaturgeschichtlicher Rezeptionsanalyse)

Mellmann, Katja

katja.mellmann@phil.uni-goettingen.de
Universität Göttingen, Deutschland

Du, Keli

keli.du@stud-mail.uni-wuerzburg.de
Universität Göttingen, Deutschland

Ausgangspunkt

Die fortschreitende Retrodigitalisierung von Kulturzeitschriften und anderen Publikationsformen mit literaturkritischen Inhalten eröffnet der literaturgeschichtlichen Rezeptionsanalyse die Möglichkeit, mit historisch repräsentativen Korpora zu arbeiten. Dabei stellt sich jedoch das Problem, dass insbesondere Zeitschriftendigitalisate in der Regel nicht als edierte Texte von standardisierter Qualität vorliegen, sondern mit 'schmutzigen Texten', also Texten mit fehlerhafter OCR und ohne linguistische Strukturierung gearbeitet werden muss. Wir wollen im Rahmen des Themas "Kritik der digitalen Vernunft" einen konstruktiven Umgang mit diesem Problem vorstellen. Wir unterscheiden dazu grundsätzlich zwei Zielperspektiven:

1. Korpusanalyse als Untersuchung mit validen Ergebnissen und
2. Korpusanalyse als Heuristik zur vorläufigen Trenddarstellung.

Die erste Perspektive ist auf qualitativ hochwertige Textkorpora angewiesen, um statistisch aussagekräftige Ergebnisse zu erzielen. Sie ist die Standardperspektive, wenn Korpusanalysen als Forschungsinstrument eingesetzt werden. Wir nehmen hingegen die zweite Perspektive ein: In ihr werden qualitativ minderwertige Textkorpora nicht als bedauerliche Abweichung vom eigentlich Gewünschten aufgefasst, sondern als Forschungsgegenstand eigener Art, der auch eine Methodik eigener Art erfordert. Diese Methodik hebt auf vorläufige Trenddarstellungen ab, die nicht als (noch unvollkommene) Vorstufe einer validen Korpusanalyse, sondern als eigenständige Heuristik zur Identifikation potentieller Ereignisse in einem diachronen Korpus aufgefasst werden. Die Methode soll sozusagen grobe Bewegungsprofile liefern, von denen aus anschließend wieder gezielte hermeneutische Tiefensondierungen unternommen werden

können. Digitale und hermeneutische 'Vernunft' stehen hier also in einem komplementären Verhältnis; nicht versucht wird, die eine durch die andere möglichst perfekt nachzubilden.

Beispielprojekt: Sentimentanalyse in historischer Literaturkritik

Bei den angezielten Bewegungsprofilen handelt es sich im Rahmen unseres Forschungsprojekts¹ um Zäsuren in der Bewertung literarischer Autoren. Wir untersuchen in einem Pilotprojekt die Rezeption von Literatur in literaturkritischen Zeitschriften des ausgehenden 19. Jahrhunderts mittels einer Sentimentanalyse der Textumgebung von Autorerwähnungen. An einem Testkorpus optimieren wir durch den Vergleich automatisierter mit manuellen Analysen eine an das historische Genre 'Literaturkritik um 1900' angepasste Sentimentwortliste.

1. Literaturwissenschaftliche Fragestellung

Historische Rezeptionsanalyse (Mellmann/Willand 2013) rekonstruiert die Aufnahme literarischer Werke durch das originale zeitgenössische Publikum. Dazu zählen insbesondere (a) Inhaltsverständnis, (b) ästhetische Wertung und (c) Kontextualisierung mit außerliterarischen zeitgenössischen Wissensformationen. Wir befassen uns in unserer Studie ausschließlich mit der ästhetischen Wertung (b).

Diese ist symptomatisch für einen umfassenden literaturgeschichtlichen Wandel in der zweiten Hälfte des 19. Jahrhunderts: Auf dem Übergang vom Bürgerlichen Realismus zur Klassischen Moderne verlieren ehemals reputierte Autoren noch während ihrer aktiven Schaffenszeit ihren führenden Status; neue Stile aus dem Bereich des gesamteuropäischen Naturalismus und Ästhetizismus gewinnen an Reputation. Für einzelne Autoren und insbesondere für poetologische Programmatiken wurde dies bereits vielfach gezeigt. Was noch aussteht, ist eine Einschätzung, wie repräsentativ die in Einzelstudien ermittelten Entwicklungen für die Gesamtentwicklung sind, insbesondere unter Einschluss auch der wenig erforschten nichtkanonischen Literatur. Abhilfe schaffen könnte hier die Analyse eines großen Korpus repräsentativer Literaturzeitschriften, die diachrone Wertungsprofile zu symptomatischen Autorengruppen liefert.

2. Korpusbildung

Langfristiges Ziel ist eine Analyse von sieben repräsentativen Zeitschriften über einen Erscheinungszeitraum von ca. 1860 bis 1900: "Die Grenzboten", "Die Gegenwart", "Deutsche Rundschau",

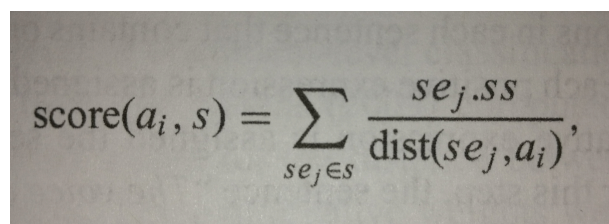
“Nord und Süd”, “Blätter für literarische Unterhaltung”, “Westermanns Monatshefte” und “Magazin für Literatur”. Als Volltext verfügbar ist derzeit nur die Zeitschrift “Die Grenzboten”. Sie dient uns als erster Anwendungsfall nach den überwachten Optimierungsläufen an einem auf der Basis einer Anthologie (Kreuzer 2006, Bd. I und II) erstellten Testkorpus.

Literaturkritisches Schrifttum im ausgehenden 19. Jahrhundert ist ein außergewöhnlich stark rhetorisches Textgenre, das durch einen hohen Grad an Ironie, Intertextualität, Euphemismus und Understatement besondere Herausforderungen an die Methode der digitalen Sentimentanalyse stellt, zumal in unstrukturierten Texten, die keine Berücksichtigung von grammatischen Komplexitäten (wie z.B. doppelter Verneinung, Konjunktiven, indirekter Rede) zulassen. Die Optimierung der Sentimentwortliste ist deshalb weniger auf eine Verfeinerung als auf eine Vergrößerung hin ausgelegt. Die Korpusanalyse soll vor allem eklatante Veränderungen identifizieren.

3. Erste Testläufe digitaler Analysen

Der Volltext des Testkorpus wurde durch NLTK Punkt Sentence Tokenizer ² tokenisiert. Mittels des Namensregisters der Anthologie wurden anschließend alle Sätze mit Erwähnung eines Autornamens extrahiert. Jeder Satz wurde in einem manuellen Rating als positiv, neutral oder negativ annotiert. Das Testkorpus umfasst 1731 1-Satz-Textsnippets. Nach der manuellen Annotation sind 505 davon positiv, 909 neutral und 317 negativ. Ausgangsbasis unserer Sentimentwortliste war die deutschsprachige Ressource “SentimentWortschatz” (Remus et al. 2010), die manuell um den im Rating auffällig gewordenen Wortschatz erweitert und um offenkundig unbrauchbare oder überflüssige Wörter gekürzt wurde. Danach erfolgte ein erster Testlauf:

Der Grundwert jedes Sentimentworts wurde mit 1 angesetzt. Die Polarität eines Satzes wurde in zwei Schritten festgelegt: Zuerst wurde das Vorkommen der positiven und negativen Sentimentwörter im Satz gezählt. Anschließend wurde die „Aggregation function“ (Abb. 1) für die Berechnung des Sentiment-Werts des Satzes verwendet: „ se_j is a sentiment expression in sentence s , $dist(se_j, a_i)$ is the word distance between aspect a_i and sentiment expression se_j in s , and $se_j.ss$ is the sentiment score of se_j “ (Liu 2015).



$$\text{score}(a_i, s) = \sum_{se_j \in s} \frac{se_j.ss}{\text{dist}(se_j, a_i)},$$

Abb. 1: Aggregation function

Ist der Sentiment-Wert größer als 0, gilt der Satz als positiv; kleiner als 0 gilt als negativ; ist der Wert gleich 0 (z.B., weil kein Sentimentwort im Satz auftaucht), gilt der Satz als neutral. Die im Satz ermittelten Sentimentwörter wurden in die Ergebnisdarstellung übernommen, um die digitale Analyse anschließend manuell überprüfen und die Sentimentwortliste optimieren zu können. Wörter, die sich als überwiegend dysfunktional erweisen, werden von der Sentimentwortliste gelöscht, fehlende Sentimentwörter werden ergänzt.

Im ersten Testlauf wurden nur ca. 47.4% der Sätze richtig erkannt (Tab. 1). Unsere Wortliste konnte vor allem die negativen Sätze schwer identifizieren. Ca. 77% der positiven Sätze wurden richtig identifiziert. Aber auch der Anteil der fälschlich als positiv klassifizierten Sätze war sehr hoch. Außerdem war der Sentiment-Wert von vielen als neutral eingestuften Sätzen nicht gleich 0.

Dimension	richtig identifizierte Sätze	Precision	Recall	F1 score	Manuelle Annotation
Positiv	387	0.4	0.77	0.52	505
Neutral	322	0.7	0.35	0.47	909
Negativ	112	0.33	0.35	0.34	317

Tab. 1: Ergebnis des ersten Testlaufs

In einem zweiten Testlauf haben wir eine automatische Klassifikation ausprobiert. Dabei wurden die Anzahl der Sentimentwörter und der Sentiment-Wert eines Satzes als Feature verwendet. Im Verhältnis 80% zu 20% wurden die Daten in einen Trainings- und einen Testdatensatz aufgespalten. Es wurde ein Support Vector Machine (SVM) Modell trainiert; die Evaluation erfolgte als 10-fache Kreuzvalidierung (Cross-Validation). Dadurch verbesserte sich das Ergebnis um ca. 10%: Die Trefferquote der Klassifikation lag bei 57% (+/- 7%).³ Wenn man eine Klassifikation nur zwischen positiven und negativen Sätzen durchführt, beträgt die Trefferquote 68% (+/- 10%).⁴

Für einen dritten Testlauf wurde die Sentimentwortliste bearbeitet und die Textsnippets wurden einem zweiten manuellen Rating mit mehr als nur 3 Kategorien unterzogen. Insbesondere sollte zwischen tatsächlich neutralen Sätzen (z.B. “X wurde 1826 in Berlin geboren.” = 0) und Sätzen mit (einander ausgleichenden) positiven und negativen Bewertungen (z.B. “Trotz dieser erheblichen Schwächen ist X ein Werk gelungen, das ...” = 0,###) unterschieden werden. Auch Problemfälle (wie z.B. erwartbare Artefakte durch Zitation oder Ironie) wurden separiert, um die Analyseergebnisse gesondert evaluieren zu können. Von den 1687 eindeutig positiven, negativen oder neutralen Sätzen wurden 65,6% richtig erkannt (Tab. 2).

Dimension	richtig identifizierte Sätze	Precision	Recall	F1 score	Manuelle Annotation
Positiv	599	0.65	0.83	0.73	718
Neutral	375	0.78	0.55	0.65	677
Negativ	133	0.47	0.46	0.46	292

Tab. 2: Ergebnis des dritten Testlaufs

In unserer Präsentation werden wir die ausführliche Ergebnisevaluation des dritten Testlaufs vorstellen und die sich stellenden Probleme im Hinblick auf die eingangs dargestellte Zielsetzung diskutieren. Außerdem soll ein erster Probelauf über das inzwischen provisorisch erstellte Satzkorpus aus den “Grenzboten” präsentiert werden, der die angezielte Methodik der diachronen Trenddarstellung illustriert.

Fußnoten

1. “Historische Rezeptionsanalyse” (gefördert von der Volkswagenstiftung), Teilprojekt “SentiLitKrit” (Göttingen, 2015-2018).
2. <http://www.nltk.org/api/nltk.tokenize.html#module-nltk.tokenize.punkt>
3. Es wurden zusätzlich Lineare SVM und Gaussian Naive Bayes ausprobiert. Die Trefferquote lag bei 0.56 (+/- 0.05) bzw. 0.56 (+/- 0.07).
4. Es wurden zusätzlich Logistic Regression, Lineare SVM und Gaussian Naive Bayes ausprobiert. Die Trefferquote lag bei 0.70 (+/- 0.10), 0.70 (+/- 0.11) bzw. 0.70 (+/- 0.11).

Bibliographie

Kreuzer, Helmut (Hg.) (2006): Deutschsprachige Literaturkritik 1870-1914. Eine Dokumentation. Unter Mitarbeit von Doris Rosenstein. 4 Bde. Frankfurt am Main: Lang.

Liu, Bing (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press.

Mellmann, Katja / Willand, Marcus (2013): Historische Rezeptionsanalyse. Zur Empirisierung von Textbedeutungen. In: P. Ajouri, K. Mellmann & C. Rauen (Hg.): Empirie in der Literaturwissenschaft. Münster: Mentis, S. 263–281.

Remus, Robert / Quasthoff, Uwe / Heyer, Gerhard (2010): SentiWS - a Publicly Available German-Language Resource for Sentiment Analysis. In: Proceedings of the 7th International Language Resources and Evaluation, pp. 1168-1171.

van Bellen, Maurits (2010): Sentiment Analysis on historical book reviews with a Bayesian Classifier. Bachelor-Arbeit. University of Amsterdam.