Erkennung und Visualisierung attribuierter Phrasen in Poetiken

Andreas Müller (1), Markus John (2), Steffen Koch (2), Thomas Ertl (2) und Jonas Kuhn (1)

(1) Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
(2) Institut für Visualisierung und Interaktive Systeme, Universität Stuttgart

Einleitung

In wissenschaftlichen Werken über Literatur (zum Beispiel Poetiken) spielen Referenzen zu Autoren, fiktiven Charakteren aus literarischen Werken und anderen Arten von Personen eine wichtige Rolle. Zum Beispiel bildet Personenerkennung eine der Grundlagen für die Erkennung der Sprecher von direkter Rede in literarischen Werken (Elson and McKeown 2010). Diese Information kann weiter benutzt werden um, zum Beispiel, soziale Netzwerke zu extrahieren (Elson et. al. 2010).

In diesem Abstrakt stellen wir eine Erweiterung der in John et. al. 2014 präsentierten Technik zum Vergleich von Textdokumenten vor. Diese Erweiterung basiert auf der Erkennung von Personennamen und Personen zugeordneten Konzepten. Ein Beispiel für ein einer Person zugeordnetes Konzept ist „Schillers Poesie“ oder „Klopstocks Messias“. Im ersten Fall wird mit der Phrase die gesamte Poesie Schillers referenziert, im zweiten Fall das konkrete Werk „Messias“ von Klopstock. Im Folgenden wird gezeigt wie man mit einem simplen, auf morphologischer Analyse, Nominalphrasenerkennung und der Erkennung von Personennamen basierendem Suchmuster solche Phrasen extrahieren kann. Nach der Extraktion werden Personennamen und Phrasen in die bereits erwähnte Technik integriert. Auf dieser Basis haben wir eine benutzerbasierte Evaluation durchgeführt, die zeigt, dass die Erweiterung der Technik durch Personennamen und Personen zugeordneten Konzepten bei literarischen Vergleichen und Analysen von Texten hilfreich ist.

Als Grundlage für unsere Untersuchung verwenden wir Poetiken aus einem Korpus von 20 Texten, die im Rahmen des Projekts ePoetics untersucht werden. Das Korpus wurde aus 1000 Poetiken ausgewählt, die von Richter (2010) analysiert wurden. Für unsere Analyse verwenden wir die vier Poetiken der Autoren Staiger, Scherer, Kleinpaul und Engel. Diese wurden von dem Experten für Literaturwissenschaft in unserem Projekt als sehr interessant für literarische Textvergleiche eingestuft.

Methode

Für die linguistische Vorverarbeitung verwenden wir die OpenNLP Tools[1] für automatische Satz- und Worterkennung. Des weiteren benutzen wir die mate tools[2] (Bohnet 2010) für Lemmatisierung, automatische morphologische Analyse und Wortartenerkennung und die StanfordCoreNLP library (Finkel et. al. 2010) mit den Modellen fürs Deutsche von Faruqui and Pado (2010) zur Erkennung von Personennamen. Für die Erkennung von Nominalphrasen benutzen wir die in MuNPEx[3] enthaltenen JAPE-Grammatiken.

Nach diesen linguistischen Vorverarbeitungsschritten suchen wir alle Personennamen, die im Genitiv auftreten. Diese signalisieren Vorkommen von den Personen zugeordneten Konzepten. Anschließend extrahieren wir für jeden Personennamen die ununterbrochene Sequenz von Nominalphrasen die dem Personennamen am nächsten ist als Konzept, das dem Personennamen zugeordnet wird. Wir extrahieren die Sequenz von Nominalphrasen statt nur der am nächsten

1 https://opennlp.apache.org/
2 https://code.google.com/p/mate-tools/
3 http://www.semanticsoftware.info/munpex

stehenden Nominalphrase um auch komplexe Nominalphrasen zu erfassen.

Integration in das System

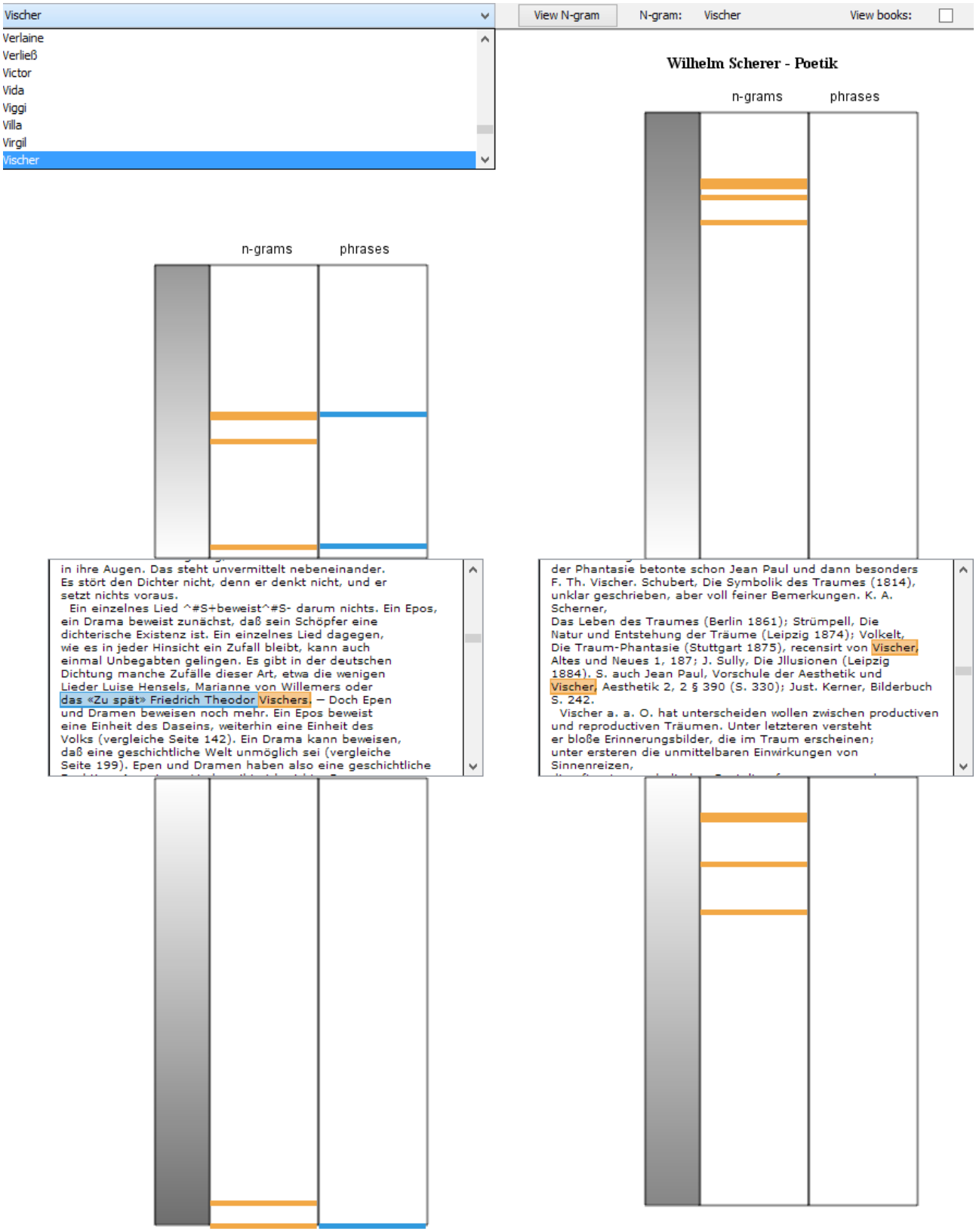

Abbildung 1. Zwei ausgewählte Dokumente werden als Band dargestellt. Die Vorkommen von Personen (Orange) und Phrasen (Blau) werden als Balken dargestellt.

In Abbildung 1 ist ein Screenshot des genannten Systems dargestellt. In der linken oberen Ecke befindet sich eine Liste von Personennamen. Nach Auswahl einer Person, werden die Vorkommen der Personen (Orange) und die der Person zugewiesenen Konzepte (Blau) in den Dokumenten als Balken dargestellt. Durch die Selektierung der Vorkommen, können Textpassagen weiterführend untersucht werden.

Literaturwissenschaftliche Evaluation

Um den Vorteil für diese Erweiterung aus literaturwissenschaftlicher Sicht zu zeigen, führten wir eine Evaluation mit einem Literaturexperten durch. Nach einer kurzen Einführung in das System, begann der Experte die 4 ausgewählten Poetiken im Hinblick auf die vorkommenden Personen zu analysieren. Es ist noch zu erwähnen, dass der Experte schon mit den Poetiken vertraut war, was die Nähe zu einem realistischen Analyse-Szenario erhöht, da eine literaturwissenschaftliche Analyse genau damit beginnt, sich mit dem Untersuchungsgegenstand vertraut zu machen, einen Text ggf. also auch mehrfach zu lesen.

Beim Durchgehen der Liste bemerkte der Analyst den Namen „Aristoteles“, der ihn interessierte. Bei der Auswahl von „Aristoteles“ fiel dem Analysten sofort auf, dass dieser Name in der Poetik von Staiger nur selten und in der Poetik von Scherer sehr häufig vorkommt, was er so nicht erwartet hatte. Solche Frequenz-basierten Eigenschaften lassen sich durch die Visualisierung sehr schnell erkennen. In der Poetik von Scherer fiel ihm weiterhin auf, dass die meisten Aristoteles zugeordneten Phrasen „Aristoteles Poetik“ oder „Aristoteles Rhetorik“ referenzierten. Dies untermauert die These, dass diese beiden Werke für Scherer eine hohe Bedeutung haben empirisch.

Eine weitere Frequenz-basierte Eigenschaft erkannte der Experte, als er „Homer“ auswählte. Im Kapitel über Epik bei Staiger (linkes Dokument mittig, siehe Abbildung 2) ist ein häufiges Vorkommen erkennbar. Dies bestärkte ihn in seiner Annahme, dass Staiger einen Hauptvertreter für jede der 3 Gattungen Epik, Lyrik und Dramatik benennt. Durch Auswahl der anderen beiden Hauptvertreter, „Goethe“ und „Schiller“, lässt sich diese Annahme empirisch untermauern, da man erkennen kann, dass „Goethe“ im Kapitel über Lyrik und „Schiller“ im Kapitel über Dramatik besonders häufig vorkommt.

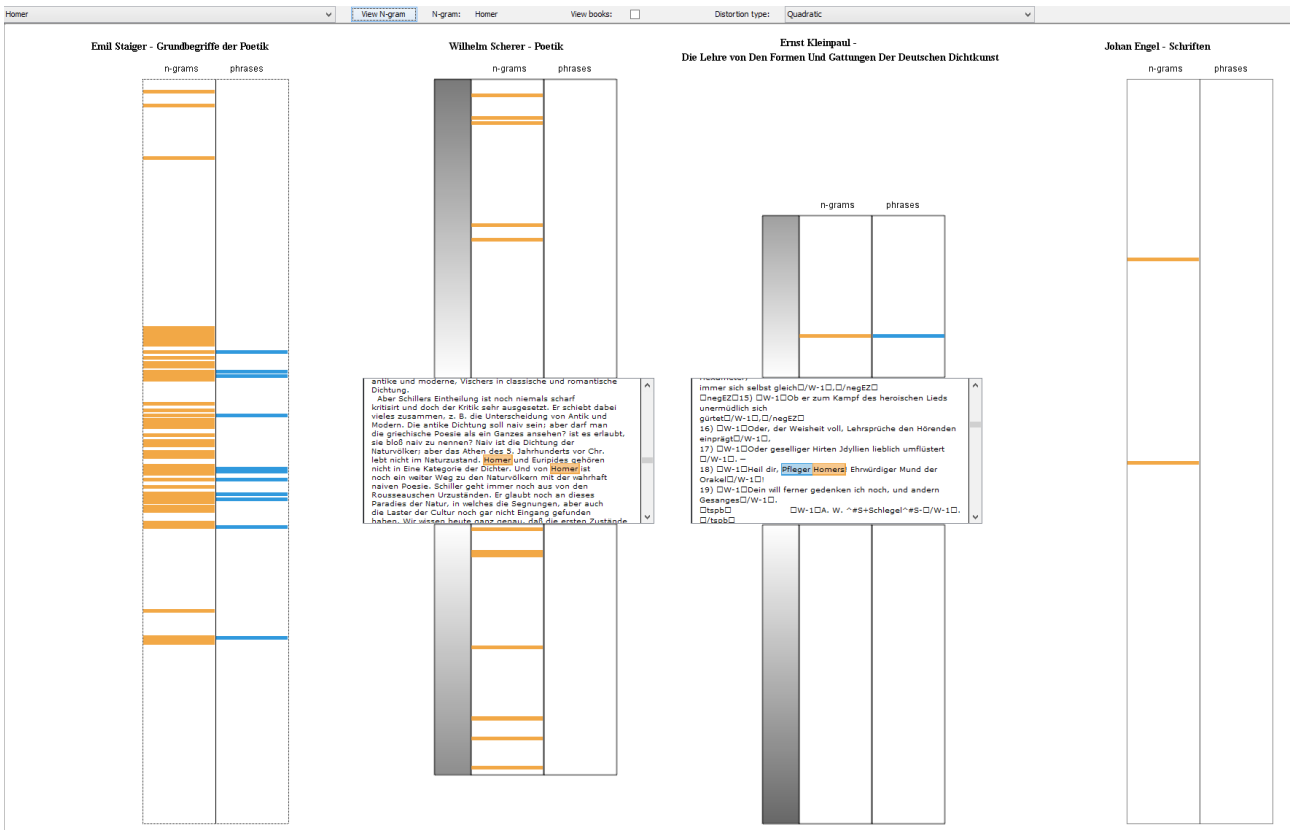

Abbildung 2: Häufiges Vorkommen von „Homer“ in dem Kapitel über Epik von Staiger (linkes Dokument mittig)

In diesem Abschnitt wurde exemplarisch gezeigt, dass die Stärken des Systems in der Übersicht über Frequenz-basierte Eigenschaften von Dokumenten und der Verwendung von Personennamen in Dokumenten liegen. Außerdem enthalten die Phrasen unter anderem Referenzen auf Personen zugewiesene Werke, wie zum Beispiel, „Aristoteles Poetik“ oder „Aristoteles Rhetorik“. Dadurch lässt sich schnell ein Überblick über Diskussionen über diese Werke und den Stellenwert der Werke in einem Dokument gewinnen. Durch die Ansicht mehrerer Dokumente, wird ein einfacher Vergleich von Textstellen über Personen oder Phrasen ermöglicht.

Technische Evaluation

Um die Präzision der Phrasenerkennung einschätzen zu können, haben wir auf Basis der Poetik von Staiger jedes erkannte attribuierte Konzept in der Poetik in eine von 4 Klassen eingeteilt, die im Folgenden aufgelistet sind. Diese Evaluation dient hauptsächlich dazu für weitere Forschungen Fehlertypen zu finden. Die 4 Klassen sind:

1. Vollständig korrekt: Personenname und attribuiertes Konzept werden korrekt erkannt

2. Teilweise korrekt weniger: Es wird mindestens ein Wort des Personennamens und des attribuierten Konzepts erkannt, aber es wird auch mindestens ein Wort des Personennamens oder des attribuierten Konzepts nicht erkannt

3. Teilweise korrekt mehr: Personenname und attribuiertes Konzept werden korrekt erkannt, aber es wird auch linguistisches Material erkannt, das weder zum Personennamen noch zum attribuierten Konzept gehört

4. Inkorrekt: Entweder wird kein Wort des Personennamens oder kein Wort des attribuierten

Konzepts erkannt. Dieser Klasse werden auch Annotationen zugewiesen bei denen es sich nicht um ein attribuiertes Konzept handelt

Diese Einteilung der erkannten Konzepte, sowie die Fehleranalyse im nächsten Abschnitt, wurden bisher nur von dem Erstautor des Abstrakts vorgenommen. Deshalb sollten die Zahlen als Schätzung der Qualität der Phrasenextraktion, nicht als definitive Evaluation angesehen werden. Eine Verifizierung der Zahlen durch einen zweiten Annotator ist geplant.

In der Poetik von Staiger erzielen wir folgende Resultate:

Vollständig korrekt: 72
Teilweise korrekt weniger: 25
Teilweise korrekt mehr: 6
Inkorrekt: 25

Es werden 56% der Instanzen komplett richtig und 20% komplett falsch erkannt. 24% der Instanzen werden nicht komplett richtig erkannt. Allerdings sind einige der inkorrekten Instanzen auf Fehler in der Erkennung von Personennamen zurückzuführen. So wird zum Beispiel „Gott“ als Personenname erkannt, was zu einem Fehler führt der mit der Erkennung der Nominalphrasen nichts zu tun hat.

Wir haben auf der Basis der 24% nicht komplett richtig erkannter Instanzen eine Fehleranalyse durchgeführt und zeigen im nächsten Abschnitt exemplarisch eine häufige Art von Fehler.

Fehleranalyse

Eine häufige Art von Fehler ist, dass oft einer komplexen Nominalphrase zugehörige Phrasen nicht erkannt werden. So wird „Goethes Forderung“ als attribuierte Phrase erkannt, nicht aber die vollständige Phrase „Goethes Forderung an ein gutes Gedicht“. Ein anderes Beispiel ist die Phrase „ein Gedicht Hebbels“. Diese Phrase wird als attribuierte Phrase erkannt, die komplette Phrase wäre aber „Ein Gedicht Hebbels, das «Lied» überschrieben ist“. Die Phrase „das Lied überschrieben ist“ beinhaltet zusätzliche Informationen, die es dem Leser erlauben zu erkennen, welches Gedicht Hebbels gemeint ist. Diese Art von Fehler lässt sich wahrscheinlich durch Einbindung eines Dependenz- oder Konstituentenparsers und darauf aufbauender Erkennung komplexerer Phrasen erkennen.

Für eine weitergehende automatische Verarbeitung der extrahierten Phrasen wären diese Fehler schwerwiegender als für die Integration in das vorgestellte Visualisierungssystem, da die Phrasen in ihrem Kontext dargestellt werden. Dadurch können Fehler schnell gefunden und vom Analysten leicht korrigiert werden. Außerdem lassen sich auch bei einer fehlerhaften automatischen Analyse zum Beispiel Frequenz-basierte Eigenschaften erkennen, solange der Fehler nicht darin besteht, dass eine Stelle markiert wird an der keine attribuierte Phrase vorliegt. Dadurch kann das System auch auf ältere Varianten von Sprachen, bei denen automatische Methoden oft nicht so gut funktionieren wie bei moderner Sprache für die sie entwickelt wurden, angewendet werden. Dies ist vor allem in literarischer Textanalyse hilfreich, da in diesem Bereich auch mit älteren Dokumenten gearbeitet wird.

Ausblick

Die Erweiterung des Ansatzes wurde im Rahmen des ePoetics Projekt entwickelt und evaluiert. Wir haben damit begonnen, das System so zu erweitern, dass die oben beschriebenen Fehler vermieden werden. Dazu verwenden wir einen Dependenzparser (Bohnet, 2010), der syntaktische Analysen

der Sätze bereitstellt. Letzten Endes sollen neben attribuierten Konzepten auch sonstige Äußerungen, Zitate und Referenzierungen von Dritten automatisch erkannt und über eine entsprechende Visualisierung zur Verfügung gestellt werden.

Referenzen:

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 89-97.

David K. Elson, Nicholas Dames, and Kathleen R. McKeown. 2010. Extracting social networks from literary fiction. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 138-147.

Elson, David K. ; McKeown, Kathleen ; Fox, Maria (Bearb.) ; Poole, David (Bearb.): Automatic Attribution of Quoted Speech in Literary Narrative.. In: AAAI : AAAI Press, 2010

M. Faruqui and S. Pado. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. Proceedings of Konvens 2010, Saarbrücken, Germany.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

Koch, Steffen; John, Markus; Wörner, Michael; Ertl, Thomas: VarifocalReader – In-Depth Visual Analysis of Large Text Documents. In: IEEE Transactions on Visualization and Computer Graphics (TVCG) (2014) (Noch nicht erschienen).

S. Richter. 2010. A History of Poetics: German Scholarly Aesthetics and Poetics in International Context, 1770- 1960. De Gruyter.