LitSOM. Kartierung russischer Gegenwartsliteratur

Gernot Howanitz, Lehrstuhl für Slavische Literaturen und Kulturen, Universität Passau
Helmut A. Mayer, Fachbereich Computerwissenschaften, Universität Salzburg

Zusammenfassung

Das literarische SOM (LitSOM) wendet selbstorganisierende Karten (Self-Organizing Maps, SOM) und Learning Vector Quantization (LVQ) auf russische Literatur an. Das SOM wird dafür eingesetzt, um eine Karte zeitgenössischer russischer Romane zu erstellen, die es Literaturwissenschaftlerinnen und Literaturwissenschaftlern erlaubt, Beziehungen zwischen den Romanen zu untersuchen. LitSOM ist eine ‚Distant reading'-Technik. Die Qualität der Karten wird sowohl subjektiv aus der Sicht der Literaturwissenschaft, als auch objektiv, d.h. in einem ‚klassischen' Problem des Textmining, nämlich der Klassifizierung von Autorinnen und Autoren, bestimmt. Um dies zu erreichen, wird der SOM-Algorithmus durch den LVQ-Algorithmus ergänzt.

1. Einleitung

1.1 Überblick

Ein Hauptziel dieses Beitrags ist die Implementierung eines Systems für computerunterstützte Textanalyse, das sogenannte *Literary SOM* (LitSOM). Darüber hinaus zeigen wir, wie Literaturwissenschaftlerinnen und Literaturwissenschaftler dieses System für ihre Forschungen einsetzen können. In unserer Beispielanwendung haben wir jeweils 15 Romane von acht verschiedenen Autorinnen und Autoren der russischen Gegenwartsliteratur kartiert, um die Nützlichkeit von quantitativen Methoden für die slavistische Literaturwissenschaft zu demonstrieren. Zwar gibt es eine lange russische Tradition quantitativer Zugänge zur Literatur, diese ist allerdings etwas in Vergessenheit geraten. So hat der bedeutende russische Mathematiker Andrej Markov 1913 ein Paper publiziert [1], in dem er das Konzept der Markovkette demonstriert. Markovketten erlauben es, Ereignisse zu modellieren, die nacheinander stattfinden. Heutzutage werden sie in verschiedensten Feldern angewandt, beispielsweise in den Wirtschaftswissenschaften oder in der Physik, und sie spielten auch eine Schlüsselrolle in Claude Shannons grundlegender Monographie zur Informationstheorie [2]. Trotz dieser vielfältigen Anwendungsmöglichkeiten hat Markov im Jahr 1913 den Versroman ‚Eugen Onegin' (1833) zu seinem Studienobjekt gemacht. Dieses Beispiel zeigt, wie stark die Verbindung zwischen Mathematik und Literatur im Russland des frühen 20. Jahrhunderts war.

1.2 Methodologie

LitSOM basiert auf sogenannten selbstorganisierenden Karten (Self-Organizing Maps, SOM) [3]. Ein SOM ist perfekt geeignet für unstrukturierte Daten und unvollständige Information, weil es hochdimensionale Probleme vereinfachen und für den Menschen leicht verständlich darstellen kann. Deshalb eignen sich SOM sehr gut für Data Mining [4]. Ein SOM kann dazu verwendet werden, eine große Anzahl von Texten zu clustern und sie auf einer zweidimensionalen Karte anzuzeigen. Das sogenannte *WEBSOM* [5] clustert beispielsweise Newsgroup-Postings nach ihrem Inhalt. Das LitSOM funktioniert ähnlich, arbeitet aber mit Romanen anstelle kurzer Nachrichten. Es erstellt eine Karte, die die Abstände zwischen einzelnen Romanen darstellt – je näher, desto ähnlicher. Dieser Text-Mining-Ansatz liefert Literaturwissenschaftlerinnen und Literaturwissenschaftlern eine automatische Visualisierung von Beziehungen zwischen unterschiedlichen Texten. Nach Franco Moretti ist LitSOM ein Werkzeug für ‚distant reading' [6], also für eine Mischung aus der klassischen literarischen Textanalyse (‚close reading') und dem Querlesen von Texten [7].

2. Vorarbeiten

2.1 Implementierung des SOM

Alle Bestandteile von LitSOM (SOM, Feature Extraction basierend auf Wortfrequenzen und eine Visualisierung mittels der Unified Distance Matrix [8]) wurden in Java implementiert. Für die Feature Extraction haben wir Sergej Sharovs Liste der 5000 häufigsten russischen Wörter verwendet [9]. Der Feature-Vektor wurde dann wie folgt zusammengestellt: Für jeden Roman wurden die Wörter aus der Sharov-Liste gezählt und jeweils durch die Gesamtanzahl der Wörter in diesem Roman dividiert. Dies gewährleistet, dass die einzelnen Feature-Vektoren untereinander vergleichbar bleiben.

2.2 Setup der Experimente

Verschiedene Längen des Feature-Vektors wurden getestet: 5, 10, 25, 30, 40, 50, 75, 100, 125, 150, 175 und 200 Features. Um den Einfluss verschiedener Wortarten auf die resultierenden Karten zu untersuchen, wurde Sharovs ursprügliche Liste modifziert; Versionen rein mit Nomen und Verben sowie eine Kontrollgruppe mit allen anderen Wörtern wurde erstellt. Aufgrund eigener Testreihen haben wir uns für ein SOM aus 108 Neuronen in einem hexagonalen 9 × 12 Gitter entschieden. Die Lernrate $\alpha(0)$ wurde auf 0.5 gesetzt und dann wie folgt verringert: $\alpha(t + 1) = \alpha(t)/(1 + \alpha(t))$. Der anfängliche Nachbaschaftsradius wurde mit 2.5 festgelegt. Nach jedem Zyklus wurde dieser Radius um 0.0005 verringert. Nach einer unüberwachten SOM-Phase mit 3000 Zyklen folgte eine überwachte LVQ-Phase mit 1000 Zyklen. Insgesamt wurden 4800 Experimente durchgeführt und ebensoviele Karten erstellt.

Tabelle 1

Rang	Feature-Vektor	Korrekt identifiziert	LVQ-Genauigkeit	1NN-Genauigkeit
1	150 Verben	103	85,83%	92,50%
2	100 Nomen	102	85,00%	95,83%
3	100 Verben	101	84,16%	91,67%
4	200 Sharov	100	83,33%	90,00%
5	175 Nomen	99	82,50%	95,00%
6	100 Sharov	98	81,67%	90,00%
6	150 Sharov	98	81,67%	90,83%
6	125 Nomen	98	81,67%	93,34%
6	125 Verben	98	81,67%	91,67%
10	75 Nomen	97	80,83%	94,16%
10	40 Nomen	97	80,83%	89,16%
10	175 Verben	97	80,83%	92,50%

3. Resultate

3.1 Klassifizierung von Autorinnen und Autoren

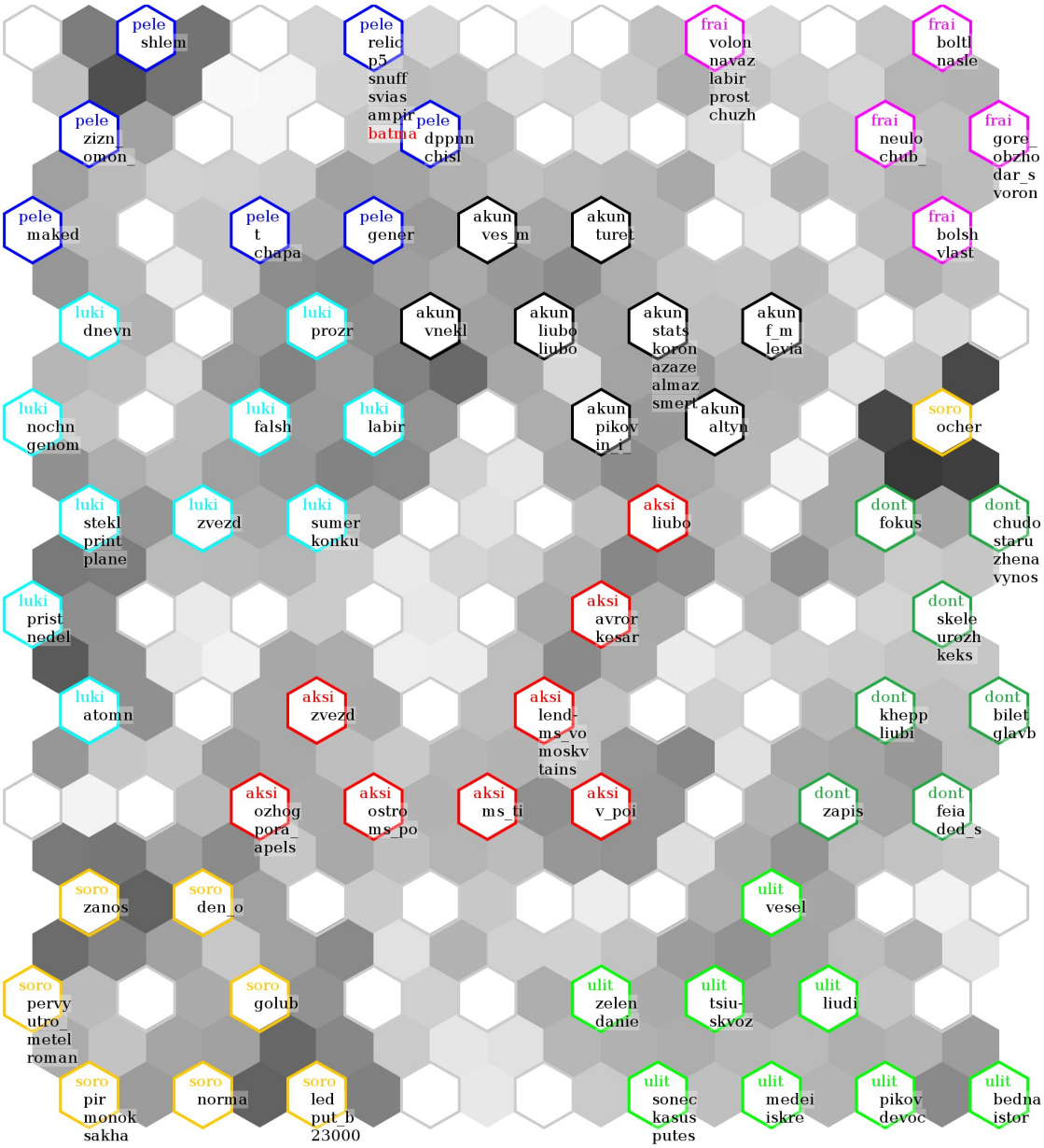
40 verschiedene Konfigurationen und Leave One Out Cross Validation (LOOCV) resultierten in einer Gesamtanzahl von 4800 verschiedenen Experimenten. Die besten Resultate dieser 4800 SOM/LVQ-Läufe sind in Tabelle 1 angeführt. Das beste Resultat – 86% richtig erkannt – wurde mit

einer Liste von den 150 häufigsten Verben als Feature-Vektor erzielt. Diese Resultate legen den Schluss nahe, dass die von LitSOM produzierten Karten die Verteilung der 120 Romane tatsächlich widerspiegelt. Mit einem 1NN-Klassifzierer, der als Kontrolle fungierte, wurde sogar eine Genauigkeit von 96% erreicht.

3.2 U-Matrix

Die Qualität der Karten kann nur subjektiv bestimmt werden. Deshalb präsentieren wir hier eine Karte samt Interpretation als Beispiel. Im Allgemeinen ist anzumerken, dass zwischen einzelnen Karten durchaus Unterschiede festzustellen waren, allerdings glichen sich die Karten trotzdem meist in ihrer grundlegenden Struktur.

Grafik 1: U-Matrix-Visualisierung des SOM für Viktor Pelevins ‚Ananaswasser für eine feine Dame‘



Grafik 1 zeigt die U-Matrix, die das LitSOM für Viktor Pelevins Roman ‚Ananaswasser für eine feine Dame‘ nach den SOM/LVQ-Läufen darstellt. Diese Karte wurde basierend auf einem Pattern-Vektor mit den 100 häufigsten Nomen erstellt. Pelevins Roman diente als unbekannter Test-Text,

d.h. nach dem Training mit den 119 anderen Romanen wurde dieser Text – korrekt – klassifiziert. Wie man sieht, weist LitSOM sehr gut auf Romane hin, die eher untypisch für die jeweilige Autorin oder den jeweiligen Autor sind. Beispiele dafür sind Vladimir Sorokins „Die Schlange“ (gekennzeichnet durch „ocher“). Auch die jeweiligen Relationen unterschiedlicher Autorinnen und Autoren zueinander sind nachvollziehbar, so liegen die beiden Fantasy-Autoren Sergej Lukjanenko und Viktor Pelevin nebeneinander. Unser letzter Test fand sozusagen unter realistischen Bedinungen statt: Pelevins neuester Roman „Batman Apollo“ wurde am 8. März 2013 publiziert, nachdem der Großteil unserer Experimente bereits abgeschlossen war, damit hat er auch nicht Eingang in das ursprüngliche Textkorpus gefunden. In Grafik 1 findet man „Batman Apollo“ („betman“ in rot) gleich neben weiteren Pelevin-Romanen jüngeren Datums, vor allem auch „Empire V“ („ampir“). Blättert man diese Romane durch, erfährt man, dass „Batman Apollo“ die Fortsetzung von „Empire V“ ist.

4. Diskussion

Die Ergebnisse der Klassifizierungsexperimente mit LVQ und 1NN sind sehr gut, vor allem in Anbetracht der Tatsache, dass literarische Texte sehr komplex sein können. Wortfrequenzen erlauben es, zwischen Romanen unterschiedlicher Autorinnen und Autoren zu differenzieren. Mit der Liste der 150 häufigsten Verben konnten 103 von 120 Romanen (86%) korrekt ihren jeweiligen Autorinnen und Autoren zugeordnet werden. Damit wurde empirisch belegt, dass LitSOM die Beziehungen zwischen Texten unterschiedlicher Autorinnen und Autoren sinnvoll darstellen kann. Die Visualisierung mittels U-Matrix, die LitSOM auch zur Verfügung stellt, erlaubt es, die Relationen zwischen unterschiedlichen Texten in einfacher Form darzustellen. Unsere subjektive Bewertung der Karten zeigte, dass die Wahl der Features großen Einfluss auf die visuelle Qualität der Karten hat. So waren die Cluster einzelner Autorinnen und Autoren bei auf der Nomen-Liste basierenden Karten am besten voneinander abgetrennt. Die Verben-Liste wiederum war für das Klassifizierungsexperiment besser geeignet, optisch waren die Karten allerdings weniger klar strukturiert. Im Allgemeinen eignen sich die Karten vor allem dazu, Texte zu finden, die für einen Autor oder eine Autorin untypisch sind bzw. die dem Stil einer anderen Autorin oder eines anderen Autors ähneln. Auch innerhalb eines Clusters lassen sich interessante Schlüsse hinsichtlich der Texte ziehen, so gibt es häufig Unterschiede zwischen noch in der Sowjet-Ära geschriebenen Texten und späteren, post-sowjetischen.
LitSOM kann in vielerlei Hinsicht verbessert werden; bei den Feature-Vektoren sind noch viele weitere Kombinationen denkbar, die durch Feature-Selection-Algorithmen bestimmt werden könnten. Weiters ist es denkbar, die visuellen Karten automatisiert durch Bildverarbeitungs-Algorithmen zu vergleichen, um eine objektivere Beurteilung zu erreichen. Auch der Einfluss der SOM-Parameter, etwa unterschiedlicher Kartengrößen, auf die visuelle Qualität der Karten ist noch nicht hinreichend untersucht. Weitere wertvolle Einblicke könnten durch Einbeziehung weiterer Texte aus unterschiedlichen literarischen Epochen gewonnen werden. Gleichzeitig würden all diese Experimente helfen, mehr Erfahrung im Umgang mit den Karten zu gewinnen. Diese Erfahrung ist sehr wichtig, denn schlussendlich kann nur ein Mensch die Karten interpretieren und als Ausgangspunkt für weitere Überlegungen nutzen – ein Prozess, der ‚distant reading‘ und ‚close reading‘ verbindet.

Quellen

1. Markov, A. 1913. Primer statisticheskogo issledovaniia nad tekstom ‚Evgeniia Onegina‘, illiustriruiushchii sviaz ispytanii v tsep’. *Izvestiia Imperatorskoi Akademii Nauk* 7.3.
2. Shannon, CE. and Weaver, W., 1949. *The Mathematical Theory of Communication.* University of Illinois Press, Illinois.
3. Kohonen, T. 2001. *Self-Organizing Maps*. Berlin.
4. Lagus, K. et al. 1999. WEBSOM for Textual Data Mining. *Artificial Intelligence Review* 13 (5/6).

http://citeseerx.ist.psu.edu/ viewdoc/summary?doi=10.1.1.12.5452 [accessed 17 March 2013].
5. Kohonen, T. et al. 2000. Self Organization of a Massive Document Collection. *IEEE Transactions on Neural Networks* 11 (3), 574–585. http://lib.tkk.fi/Diss/2000/isbn9512252600/article7. pdf [accessed 14 March 2013].
6. Moretti, F. 2000. Conjectures on World Literature. *New Left Review* 1, 54-68. http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature [accessed 30 November 2013].
7. Kirschenbaum, M. 2007. The Remaking of Reading: Data Mining and the Digital Humanities. *National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation.* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.959&rep=rep1&type=pdf [accessed 9 May 2013]
8. Kohonen, T. 2001. *Self-Organizing Maps.* Berlin, 165f.
9. Sharov, S. 2001. Chastotnyi slovar'. *RosNII II Website*. http: //www.artint.ru/projects/frqlist.php [accessed 12 March 2013].