Die automatische Erkennung des Geschlechts von Personen in deutschen Romanen anhand der Themen in ihren wörtlichen Reden mittels maschinellen Lernens

Masterpraktikum

Joachim Bergmann

Julius-Maximilians-Universität Würzburg
Institut für Informatik
Lehrstuhl für Künstliche Intelligenz

Zusammenfassung. Ist es möglich Rückschlüsse auf das Geschlecht einer Person allein anhand ihrer Wortwahl zu ziehen? Diese Arbeit zeigt einen ersten Ansatz, Protagonisten aus Romanen anhand ihrer direkten Reden einem Geschlecht zuzuordnen. Hierbei soll ein Themenmodell und klassische maschinelle Lernmethoden wie SVM und Maximum Entropy Verwendung finden.

Inhaltsverzeichnis

1 Einführung

Diese Arbeit widmet sich der Frage, ob sich das Geschlecht einer Romanfigur allein anhand ihrer direkte Rede ausmachen lässt. Hierbei sollen nur die Worte der Aussagen und eine Zuordnung der Reden zur aussprechenden Person verwendet werden. Um vom reinen Wortlaut der Aussagen Schlussfolgerungen zu ziehen findet ein Themenmodell Einsatz. Das maschinelle Lernen zur Geschlechtsidentifikation wird mit Vektoren, welche die Themen der Aussage widerspiegeln, durchgeführt.

Nachfolgend wird die aktuelle Forschung auf diesem Themengebiet betrachtet. Es schließt sich eine Aufstellung der verwendeten Techniken und den Zusammenhang der einzelnen Komponenten an. Im folgenden Kapitel werden die Ergebnisse der Evaluation betrachtet und Möglichkeiten zur tiefer gehenden Betrachtung dargestellt. Abschließend werden die Ergebnisse aufgegriffen und weiterführende Fragen aufgezeigt.

2 Stand der Forschung

Bamman et al. forscht an der automatischen Charaktererkennung in Filmen und englischen Romanen [1,2]. Hier fließen jeweils zahlreiche weitere Metainformationen in die Erkennung ein. So verwendet Bamman et al. in seiner Arbeit zur Charaktererkennung in Filmen zum Beispiel die Wikipediaartikel der Filmbeschreibung [1]. In englischen Romanen sind für die Erkennung Verben und deren Charakterisierung ein Kernmerkmal [2]. Die Analyse von direkten Reden

hat Celikyilmaz et al. mit der Entwicklung des „Actor-Topic Model“[1] (ACTM) vorangetrieben[5]. Dieses Modell beschreibt die Gesprächsthemen und die beteiligten Personen. Eine Charakterisierung der Akteure selbst findet nicht statt.

3 Verwendete Techniken

Dieses Kapitel behandelt die verwendeten Techniken zur Erkennung eines Geschlechts anhand ihrer Aussage. Zuerst wird die nicht überwachte Themengenerierung betrachtet, danach die verwendeten Klassifizierungen, die in die Kategorie der überwachten Lernalgorithmen einzuordnen sind. Einen Überblick über die Bestandteile und Verarbeitungsschritte zeigt Abbildung 1. Als Datenausgangsbasis stehen Romane im reinen Textformat ohne weitere Metainformationen und direkte Reden, die einzelnen Personen zugeordnet sind, zur Verfügung. Die Implementierung fußt in weiten Teilen auf dem Mallet Toolkit von Andrew McCallum [10].

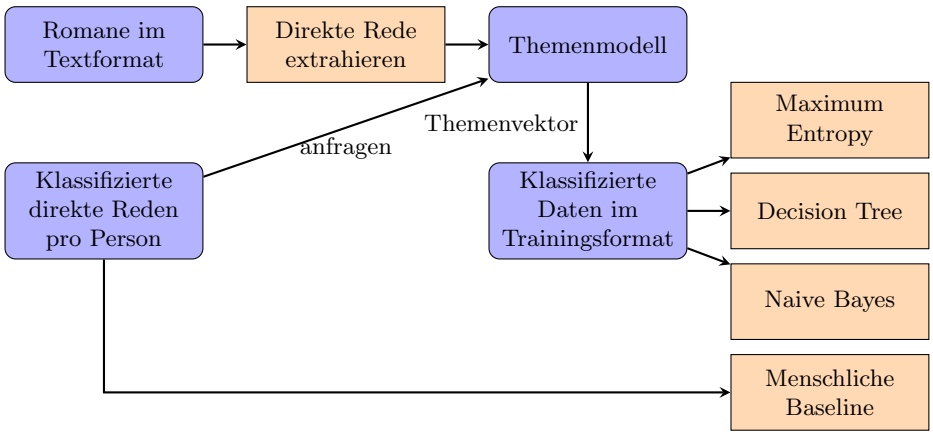

Abb. 1: Übersicht über die Daten und deren Verarbeitungsschritte.

3.1 Themengenerierung

Die Themen werden mittels einer LDA[2] aus den direkten Reden berechnet, dieses Verfahren wurde von Blei et al. vorgestellt[4]. Zunächst müssen jeweils die wörtlichen Reden extrahiert werden.

Jeder Roman liegt als eigene txt-Datei vor, wobei der komplette Text ohne weitere Metadaten gegeben ist. Direkte Reden sind in den Romanen jeweils durch

[1] Sprecher-Themen Modell
[2] Latent dirichlet allocation

die Zeichen „≫“ und „≪“ eingerahmt. Diese direkten Reden werden über einen einfachen regulären Ausdruck `≫.*?≪` gesucht und in dem Mallet kompatiblen Format auf der Festplatte gespeichert. Das Mallet Toolkit erwartet pro Zeile einen Datenpunkt in folgendem Format:

```
NAME        LABEL        DATA
```

Der Name wird einfach generisch nach oben gezählt. Da das Label für die LDA irrelevant ist, wird hier immer ein „X“ eingefügt. Der Platzhalter „DATA“ wird stets durch die jeweilige wörtliche Rede ersetzt.

Zur eigentlichen Themengenerierung wird eine Pipe, bestehend aus Malletelementen, aufgebaut. Bestandteile dieser Pipe sind: Umwandlung in Kleinbuchstaben und Stopwortentfernung mittels einer deutschen Stopwortliste. Die LDA wird mit den Werten $alpha_t = 1.0, beta_w = 0.01$ ausgeführt, die Anzahl der Themen variiert von 75 bis 200 in 25er Schritten. Jedes Themenmodell wird mit 2000 Iterationen berechnet.

3.2 Personen identifizieren

Für die Zuordnung der Aussagen zu einer Person werden Methoden, wie Krug et al. sie beschrieben hat, verwendet [8]. Zuerst werden in einem Roman die einzelnen Figuren und deren Referenzen erkannt [7]. Anschließend werden mit einer Koreferenzauflösung die direkten Reden den erkannten Figuren zugeordnet [9].

Das Verfahren erzeugt pro erkannter Person eine Datei. In der ersten Zeile sind alle erkannten Zuordnungen enthalten, Namen und die Referenzen. In den folgenden Zeilen sind die zugeordneten direkten Reden, eine Zeile pro Aussage, gegeben.

3.3 Vorbereiten der Trainingsdaten

Als Grundlage dienen von Hand klassifizierte Daten. Die direkten Reden werden als ihr Themenvektor mit ihrer Klassifizierung abgespeichert und im Mallet kompatiblen Format gespeichert.

Hierbei werden sämtlichen Aussagen einer Person einem Geschlecht zugeordnet, wodurch viele direkte Reden auf einmal klassifiziert sind. Diese Dateien werden in je einen Ordner „m“ und „f“ einsortiert. Für jede dieser klassifizierten Personen wird für jede direkte Rede mittels des Themenmodells der Vektor berechnet. Alle Vektoren werden mit ihrer Klassifikation als Label in einer Datei gespeichert, wobei die Themenzugehörigkeit auf Prozent ohne Nachkommastellen gerundet wird. Dieses Vorgehen verdeutlicht der Code 1.

3.4 Klassifikationsmodelle

Zum Vergleich werden drei verschiedene Klassifikationsmodelle betrachtet. Maximum Entropy, Decision Tree und Naive Bayes, alle drei sind im Mallet Toolkit[10],

Code 1 Dieser Pseudocode zeigt den Erstellungsablauf der Trainingsdatei für die weiblichen Personen im Malletformat.

```
for each File in 'f'
  for each Line, exclusive first line
    //getTopicVector(String line) - calculate the topicdistribution
    //for the parameter String line
    Vector topicVector = TopicModel.getTopicVector(line);

    for i=0 to topicVector.length
      trainfile.append("Topic"+i+"="+round(topicVector[i])+" ");
    trainfile.append("|f"); //the classification label
```

inklusive Trainer, vorhanden. Sie werden über die jeweils gleichen Schnittstellen angesprochen. Eine genauere Betrachtung der Klassifikationsmodellen ist in der Literatur zu finden [3], [6] und [11].

3.5 Generierung der menschlichen Baseline

Zur besseren Einschätzung der Effektivität des Maschinellen Lernens wird ein Vergleich benötigt. Hierzu dienen die Ergebnisse einer Klassifizierung durch einen Menschen, die mit einem kleinen Konsolenprogramm erzeugt werden. Dieses gibt die klassifizierten direkten Reden in einer zufälligen Reihenfolge aus und wartet auf die Klassifizierung durch den Bediener. Die Eingaben werden mit der richtigen Zuordnung überprüft. Bei der Beendung des Programms werden die ermittelten Werte ausgegeben.

4 Evaluierung

Zur Betrachtung und Analyse der Qualität der Klassifizierungen wird ein Vergleich gezogen zur menschlichen Erkennung des Geschlechts. Die möglichen Fehler und Grenzen der Optimierungen werden im Anschluss aufgezeigt.

4.1 Trainingsdaten

Zur Themenerkennung werden insgesamt 452 Romane aus dem deutschsprachigen Raum herangezogen. Aus diesen Romanen werden mittels des regulären Ausdrucks 548.788 direkte Reden extrahiert. Diese bilden insgesamt den Korpus das Themenmodells mit 75 bis 200 Themen in 25er Schritten gestaffelt.

Zur Klassifizierung werden jeweils 34 Männer und Frauen zufällig durch einen Menschen ausgewählt, bevorzugt Personen mit vielen wörtlichen Reden. Diese kommen zusammen auf 8133 direkte Reden, wobei 36% der Reden von Frauen und 64% der Reden von Männern sind. Für die menschliche Baseline werden insgesamt 200 direkte Reden klassifiziert.

4.2 Ergebnisse

Um eine Vergleichbarkeit der Klassifizierungen herzustellen wird ein 10-Fold[3] durchgeführt. Die menschliche und „Alle sind Männer“ Baselines werden nicht über Kreuz validiert. In der Tabelle 1 sind die höchsten erreichten Werte der Genauigkeit mit Anzahl der Themen angegeben. In Abbildung 2 ist die Genauigkeit[4] der Klassifizierungen im Verhältnis zur Anzahl der Themen im Themenmodell aufgetragen.

Tabelle 1: Die maximalen Genauigkeiten, die bei den Klassifikationen erreicht werden, mit der Anzahl der Themen bei der die Werte erreicht werden.

Klassifikation	Anzahl der Themen	Genauigkeit
Menschliche Baseline		51,5%
„Alle sind Männer“		64,3%
Maximum Entropy	150	76,5%
Naive Bayes	150	61,5%
Decision Tree	75	64,5%

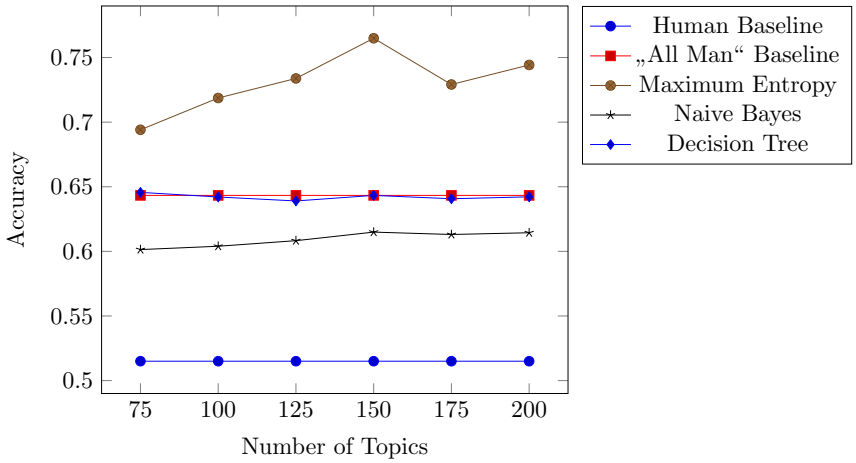

Abb. 2: Zeigt die Genauigkeit der Klassifikationen. „Human Baseline“ und „All Man Baseline“ sind unabhängig von der Anzahl der Themen.

[3] Kreuzvalidierung mit 10 Teilmengen

[4] $\frac{Korrekt\ Klassifizierte}{Anzahl\ der\ Klassifizierungen}$

In den Abbildungen 3 und 4 ist der Verlauf des F1-Maßes für die Erkennung der Frauen und Männer aufgezeigt. Die menschliche Baseline für die Frauen liegt bei 36,1%, der maximale Wert wird von Maximum Entropy bei 150 Themen und einer Genauigkeit von 53,1% erreicht. Bei den Männern ist die Baseline 60,6% und der maximale Wert, mit der Klassifikation Decision Tree bei 200 Themen, 99,1%.

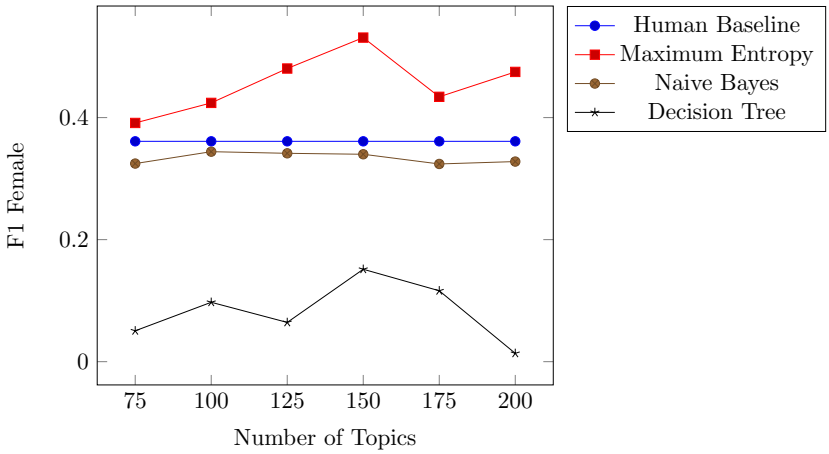

Abb. 3: Zeigt das F1-Maß der Klassifikationen zur Erkennung der Frauen. „Human Baseline“ ist unabhängig von der Anzahl der Themen.

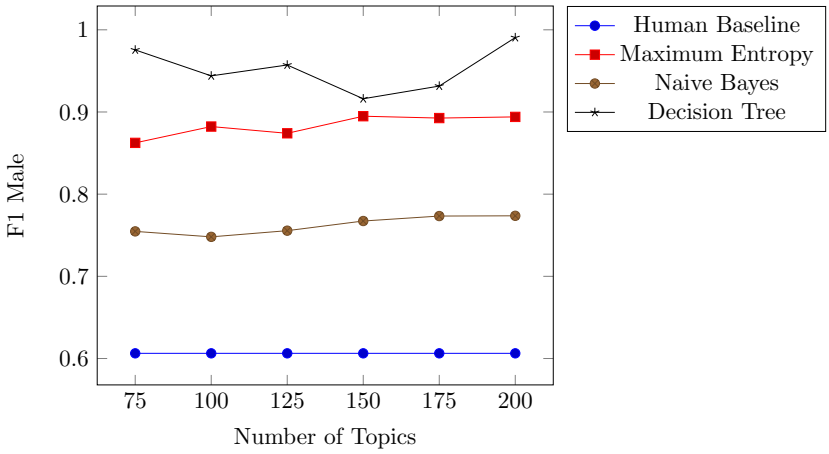

Abb. 4: Zeigt das F1-Maß der Klassifikationen zur Erkennung der Männer. „Human Baseline“ ist unabhängig von der Anzahl der Themen.

4.3 Fehlersuche

Bei der automatischen Erkennung wird eine maximale Genauigkeit von 76,5% erreicht. Nun werden Fehler ermittelt und analysiert an welchen Stellen noch Verbesserungspotential besteht. Die einzelnen Teilschritte vor den Klassifizierungsalgorithmen werden nachfolgend betrachtet und mögliche Fehler aufgezeigt.

Themengenerierung Als ersten Schritt wird die Extraktion der direkten Reden und der daraus folgenden Themen betrachtet.

Bei einer händischen Einsicht der Romantexte fällt auf, dass nicht alle direkten Reden durch den Ausdruck ≫`.*?`≪ identifizierbar sind. So sind in manchen Romanen die direkten Reden durch die Zeichen `> <` gekennzeichnet. In anderen Werken ist die wörtliche Rede durch kein seperates Zeichen vom weiteren Text hervorgehoben. Die direkte Rede lässt sich in diesem Fall nur unter Berücksichtigung von einleitenden Verben extrahieren. Eine zusätzliche Verwendung des Ausdrucks `>.*?<` ist leicht umsetzbar. Zur Identifizierung der anderen wörtlichen Reden bedarf es einer umfangreicheren syntaktischen Analyse des Textes.

Die generierten Themen beinhalten, unabhängig von der Anzahl der Themen, häufig Eigenname von Personen. Diese Namen sind nur im Kontext eines Werkes stellvertretend für das Thema und verfälschen die Zuordnung für andere Romane. So zeigt unten stehendes Beispiel als Thema ein Freudenhaus, in dem offensichtlich eine Frau mit Namen Justine tätig ist. Ihr Name ist das häufigste Wort dieses Themas. Wird in einem anderen Werk nun der Name Justine verwendet, so bildet der Name immer auf dieses Thema ab. Diese Name-Themenverknüpfung ist jedoch irreführend. Eine Identifizierung von Eigennamen, um diese gezielt zu ignorieren, könnte die Themen und deren Zuordnung verbessern.

```
justine; frau; glied; mädchen; rose; hure; hinten; schönen; hintern;
venus; deinen; mund; tochter; welch; göttin; vergnügen; jungfrau;...
```

Klassifizierte Trainingsdaten Durch die manuelle Klassifikation bei der Erstellung der Baseline sind offensichtlich falsche Trainingsdaten erkennbar.

Die Romane beinhalten auch Ein-Wortsätze, wie zum Beispiel „Ja“ oder „Nein“. Durch eine längere Antwort wie „Ja, genau.“ wird der Satz nicht aussagekräftiger. Diese Beispiele sind nicht geschlechtsspezifisch. Weder im Training noch in der anschließenden Klassifizierung lassen sich daraus relevante Ergebnisse ableiten. Ebenso ist auch der Themenvektor für diese Sätze nicht spezifisch. Eine einfache Nichtbeachtung von kurzen Sätzen schafft keine Abhilfe, da Sätze wie „Ich stille mein Kind“ stark geschlechtsspezifisch sind und wahrscheinlich ein Thema identifizieren.

Durch die Verwendung von Koreferenz bei der zugeordneten direkten Rede sind die Trainingsdaten nicht fehlerfrei. So sind deutlich geschlechtsspezifische Aussagen falsch klassifiziert. Der Satz ‘Ich stille mein Kind“ ist als männlich klassifiziert und „Ich, Sam Parker“ als weiblich. Ebenso gibt es Datensätze die bei der Klassifikation für das Training nicht eindeutig waren, da sich die Namenszuordnungen widersprochen haben. Damit sind die zugeordneten direkten

Reden nicht fehlerfrei und es ist zu überlegen, ob sie als automatisch erzeugte Daten verwendbar sind. Die Alternative wäre ein manuelles Lesen der Texte und klassifizieren der Aussagen, welches aufgrund der großen Anzahl an benötigten Daten nicht in Betracht kommt.

5 Fazit

Eine Genauigkeit von 76,5% in der maschinellen Klassifikation ist nur eine Verbesserung um 10% zur „Alle sind Männer"-Klassifikation, dies birgt noch Verbesserungspotential. Im Vergleich zur menschlichen Baseline 51,5% wurde eine Verbesserung auf das 1,5-Fache geschafft. Dieses Kapitel greift zuerst kurz die erkannten Fehler mit ihren Schlussfolgerungen auf. Danach werden weitere Aspekte des Versuches und grundlegendere Fragestellungen zum Verfahren dargestellt.

5.1 Schlussfolgerungen aus der Evaluierung

Aus dem Kapitel 4.3 „Fehlersuche" lassen sich direkt Anregungen für eine weiterführende Forschung entnehmen.

Zur Extraktion der direkten Reden sollte auch der regulären Ausdruck `>.*<` verwendet und mittels syntaktischer Analysen weitere direkte Reden identifiziert werden. Eine Verbesserung des Themenmodells wird auch erwartet, wenn gezielt Eigennamen erkannt und aus der weiteren Verarbeitung ausgeschlossen werden. Dies gilt für die Themenmodellgenerierung und für die Klassifikationen.

5.2 Weiterführende Fragestellungen

Im Verlauf der Auswertung tauchen noch weitere Fragen auf. Diese wurden bisher nicht vertieft betrachtet und werden an dieser Stelle nur vorgestellt.

Es gibt in den Romanen teilweise wörtliche Reden einzelner Personen mit sehr langen Passagen. So können Aussagen mit 40 Sätzen und 770 Wörtern identifiziert werden. Diese direkten Reden umfassen wahrscheinlich mehr als ein Thema und verfälschen somit das Themenmodell. Eine detaillierte Analyse der langen Aussagen gibt Aufschluss darüber, ob eine gesonderte Behandlung erforderlich wäre und wie diese unter Umständen aussähe. Hierzu ist interessant, wie eine direkte Rede als „zu lang" definiert werden soll.

Bei diesem Ansatz wurden die wörtlichen Reden immer nur als einzelne Elemente betrachtet. Eine direkte Rede wurde klassifiziert und auf deren Richtigkeit überprüft. Unter Umständen erschließt sich das Geschlecht einer Person nur aus der Menge aller ihrer Aussagen. Mit einem größeren Trainingskorpus wäre dies umsetzbar. Pro Person müssten alle ihre Aussagen klassifiziert und das Geschlecht über den Mittelwert bestimmt werden. Erst danach sollten die Ergebnisse in die Statistik einfließen. Der K-Fold müsste dementsprechend nicht mit Aussagen sondern mit Personen arbeiten, was eine deutlich aufwendigere Datenbehandlung erfordert.

Lässt sich das Geschlecht einer Romanfigur an deren Gesprächsinhalt erkennen? Aus dieser Frage lässt sich weiterführend formulieren, ob es eine signifikante Abweichung zwischen weiblichen und männlichen Autoren gibt. Eine differenzierte Betrachtung wäre mit einem entsprechend aufbereiteten Trainingskorpus leicht umsetzbar. Ebenso ist eine Unterscheidung in der Zielgruppe interessant, lesen zum Beispiel Frauen primär Bücher wo weibliche Protagonisten besser erkennbar sind? Hierfür müsste wahrscheinlich ein modernerer Korpus verwendet werden, um eine Zielgruppenunterscheidung zu ermöglichen.

Literatur

1. Bamman, D., O'Connor, B., Smith, N.A.: Learning latent personas of film characters. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). p. 352 (2013)
2. Bamman, D., Underwood, T., Smith, N.A.: A bayesian mixed effects model of literary character. In: ACL (1). pp. 370–379 (2014)
3. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational linguistics 22(1), 39–71 (1996)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research 3, 993–1022 (2003)
5. Celikyilmaz, A., Hakkani-Tur, D., He, H., Kondrak, G., Barbosa, D.: The actortopic model for extracting social networks in literary narrative. In: NIPS Workshop: Machine Learning for Social Computing (2010)
6. Friedl, M.A., Brodley, C.E.: Decision tree classification of land cover from remotely sensed data. Remote sensing of environment 61(3), 399–409 (1997)
7. Jannidis, F., Krug, M., Reger, I., Toepfer, M., Weimer, L., Puppe, F.: Automatische erkennung von figuren in deutschsprachigen romanen. In: Conference Presentation at" Digital Humanities im deutschsprachigen Raum (2015)
8. Krug, M., Jannidis, F., Reger, I., Macharowsky, L., Weimer, L., Puppe, F.: Attribuierung direkter reden in deutschen romanen. DHd 2016 p. 181 (2016)
9. Krug, M., Puppe, F., Jannidis, F., Macharowsky, L., Reger, I., Weimer, L.: Rulebased coreference resolution in german historic novels. on Computational Linguistics for Literature p. 98 (2015)
10. McCallum, A.K.: Mallet: A machine learning for language toolkit (2002), http://mallet.cs.umass.edu
11. Rish, I.: An empirical study of the naive bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. vol. 3, pp. 41–46. IBM New York (2001)