eAQUA Wissensdatenbank

Neue Methoden in den Geistenwissenschaften

Die zunehmende Einspeisung von kompletten Textsammlungen in elektronische Systeme hat am Ende des 20. und Beginn des 21. Jahrhundert zu einer neuen Situation in den Wissenschaften geführt. In diesem Zusammenhang ist häufig von **Text Mining** die Rede. Es handelt sich hierbei um einen Oberbegriff, der unterschiedliche statistische und linguistische Verfahren subsimiert.

In eAQUA, ursprünglich ein BMBF-gefördertes Projekt im Programm „Wechselwirkungen zwischen Geistes- und Naturwissenschaften“, sind einige dieser Verfahren mit Blick auf die historischen Sprachen Griechisch und Latein untersucht und weiterentwickelt worden. Im Ergebnis präsentiert sich eAQUA als Portal, in dem entwickelte Tools hinsichtlich ihrer Gebrauchsfreudigkeit zu abgeschlossenen Korpora, die sich dieser historischen Sprachen bedienen, ausprobiert werden können. Zwei von diesen Tools, die sogenannten Kookkurrenz- und Zitationsanalyse, sollen nachfolgend näher erläutert werden.

Verarbeitung von Sprache

Für die Gewinnung strukturierter Informationen aus Texten kommen, je nach Anwendungsfall, verschiedene Sprachtechnologie-Komponenten zum Einsatz. Bei der Verarbeitung antiker Texte ergeben sich, beispielsweise durch das Fehlen von sogenannten Metadaten, ein paar Besonderheiten, so dass nicht alle Komponenten berücksichtigt werden. Nachfolgend soll ein grober Umriss der zum Einsatz kommenden Sprachtechnologie gegeben werden.

Grundsätzlich wird innerhalb von Data-Mining bei der Verarbeitung von Sprache von drei Bereichen gesprochen:

- domänenspezifische Verarbeitung
- dokumentspezifische Verarbeitung
- sprachspezifische Verarbeitung

Bei dieser Aufzählung handelt es sich um eine thematische, nicht chronologische.

Domänenspezifische Verarbeitung

Teilaufgabe	Erläuterung
Eigennamenextraktion	Erkennung von spezifischen Entitäten; meist auf der Basis manuell annotierter Datensätze. Hierbei sind nur die für die Domäne (das Korpus) typischen gemeint. [1)]
Stopwortliste erstellen	Eine Stopwortliste ist eine Liste mit Begriffen, die bei der späteren Verarbeitung ausgenommen werden sollen. [2)]
Topic-Modellierung	Automatische Zuordnung von Begriffen zu Themen auf Basis von Worteigenschaften und Kontextinformationen.
Faktenextraktion	Vorher definierte Arten von Informationen werden durch die Verarbeitung modelliert. Viele Verfahren nutzen dafür die Abfolge unterschiedlicher Wörter in einem Satz. [3)]
Relationsextraktion	Erkennung von Beziehungen zwischen Entitäten in einem Text.

Dokumentspezifische Verarbeitung

Teilaufgabe	Erläuterung
Metadaten erfassen	Metadaten, im Falle der Korpusanalyse z.B. Enstehungsort, Entstehungszeit, Autorenschaft, Editor, Editionszeit usw., sind bei der Textanalyse wertvolle Informationsquellen, um beispielsweise die Auswahl der zu verarbeitenden Daten einzugrenzen.

Teilaufgabe	Erläuterung
Bereinigung und Normalisierung	Abhängig davon, wie die Daten erfasst wurden, müssen sie vor der Analyse von allen irrelevanten Informationen, wie z.B. die für Auszeichnungssprachen üblichen Markup Tags, bereinigt werden. Eventuell abweichende Zeichenkodierungen, wie z.B. transkripierter altgriechischer Beta-Code, müssen vor der Verarbeitung in eine einheitliche Zeichenkodierung konvertiert werden.

Sprachspezifische Verarbeitung

Teilaufgabe	Erläuterung
Spracherkennung	Die verwendeten Sprachen werden ermittelt. [4]
Segmentierung	Strukturiert den Text in einzelne Teile, die separat untersucht werden können. Üblich ist die Segmentierung in Sätze anhand der Satzzeichen.
Tokenisierung	Segmentiert auf der Basis der Wortebene in einzelne Teile (Token), indem beispielsweise das Leerzeichen als Wortgrenze aufgefasst wird.
Wortstammreduktion	Die Wörter werden auf ihren Wortstamm zurückgeführt, um bei einer späteren Suche auch Flexionen zu finden.
Lemmatisierung	Die Grundform eines Wortes (Lemma) wird gebildet.
Part-of-Speech Tagging	Zuordnung von Wörtern und Satzzeichen in Wortarten.
Parsing	Der Text wird in eine neue syntaktische Struktur überführt. Dabei ist für den Parser ein Token die atomare Eingabeeinheit.
Koreferenz (Referenzidentität) auflösen	Eine Koreferenz liegt vor, wenn sich innerhalb einer Äußerung zwei sprachliche Ausdrücke auf das selbe linguistische Objekt beziehen, beispielsweise mittels Verwendung von Pronomen.
Eigennamenextraktion	Bei der Eigennamenerkennung, auch Named Entity Recognition (NER), werden die Begriffe eines Textes bestimmten Typen (z.B. Ort oder Person) zugeordnet.

Kookkurrenz-Berechnung

Kookkurrenz bezeichnet in der Linguistik allgemein das gemeinsame Auftreten zweier lexikalischer Einheiten innerhalb eines übergeordneten Segmentes. Treten beispielsweise zwei Terme häufig gemeinsam in einem Satz auf, besteht eine berechtigte Annahme eines Abhängigkeitsverhältnisses, ob semantischer oder auch grammatikalischer Natur. Über statistische Berechnungen werden Maße für die vermutete Abhängigkeit ermittelt. Dazu müssen mehrere Voraussetzungen erfüllt sein:

- Es muss ein Gesamtkorpus definiert sein, in dem das Auftreten von Einheiten, also z.B. Wörten, gezählt werden kann. Diese statistischen Kenngrößen bilden die Berechnungsgrundlage.
- Das Korpus muss segmentiert werden. Für Nicht-Nachbarschaftskookkurrenzen, also Kookkurrenzen, die keine direkten linken oder rechten Nachbarn sind, werden maximal definierte Distanzen benötigt, ansonsten würde ja jedes Wort mit jedem Wort in einem Text verbunden sein. In eAQUA hat man sich für die lexikalische Einheit Satz entschieden. Neben den nachbarschaftlichen Kookkurrenzen werden demzufolge Satzkookkurrenzen angeben.
- Das Korpus muss sehr gross sein oder nur die häufigsten Wörter in die Berechnung einbeziehen. [5]

In eAQUA wurden die **Nachbarschafts- und Satzkookkurrenzen** vor allem mit Wahrscheinlichkeitsfunktionen und dem Signifikanzmaß Log-Likelihood berechnet. Die ermittelten Werte sind lediglich in ihrer relativen Ordnung aussagekräftig, im Gegensatz beispielsweise zum Dice- oder Jaccard-Koeffizienten, die immer einen absoluten Wert zwischen 0 und 1 liefern. Grundsätzlich gilt hier bei dem so bezeichneten „lgl“-Wert: je größer der Wert um so wahrscheinlicher ein Zusammenhang, wobei es aufgrund des Algorithmus auch zu negativen Werten kommen kann.

Zitations-Analyse

Die Zitations-Analyse beschäftigt sich als Teilgebiet der Bibliometrie mit der qualitativen Untersuchung von zitierten und zitierenden Arbeiten. Die Ergebnisse werden in einem Zitationsgraphen visuell aufbereitet. Daran

lassen sich verschiedene Regelmäßigkeiten und Strukturen eines Autors bzw. einer Autorengruppe ablesen. Falls die entsprechenden Meta-Daten vorhanden sind [6], können die Darstellungen durch eigene Suchfilter eingrenzt werden.

Die Zitations-Analyse wird anhand von String-Matching-Algorithmen vorgenommen. Zeichenkettenalgorithmen suchen nach exakten Übereinstimmungen eines Musters in einem Text unter Definition von Toleranzkriterien. Diese Kriterien wurden in der Zitations-Analyse von eAQUA wie folgt festgelegt.

Reduziert um alle Satzzeichen und einer Liste der häufig benutzten Wörter [7] wird das Korpus in eine Folge von fünf aufeinander folgenden Terme zerlegt und mithilfe eines sogenannten naiven Algorithmus auf exakte Übereinstimmungen (matches) im Restkorpus hin untersucht. Das Restkorpus ist nicht reduziert durch Berücksichtigung von Metadaten, wie beispielsweise den Entstehungszeitpunkt. Eine Eigenheit dieser Vorgehensweise ist, dass bei einigen Autoren Selbstzitate gefunden werden, also Stellen, an denen sie sich offensichtlich wiederholen. Eine andere, dass ein Zitat aus mehreren Einträgen besteht kann [8] und erst über die Sortierfunktion als Ganzes erkennbar wird.

Die Parallelstellen werden schlussendlich unter Verwendung der Editierdistanz mit einem Similaritätswert belegt, der zwischen 0 = nicht identisch und 1 = vollständig identisch liegt. Berechnet wird nach einem Algorithmus **Similar-Text**, der bei Oliver [9] mittels eines Pseudo-Codes beschrieben ist.

$$sim = \frac{n_{ab} \times 2}{n_a + n_b}$$

wobei n_a und n_b jweils die Zeichenkettenlänge der zu vergleichenden Teiletexte und n_{ab} die Anzahl der identischen Zeichen sind, also die Differenz zur **Levensthein-Distanz** [10].

$$sim = \frac{\left(\max\left(n_a, n_b\right) - lev\left(a,b\right) \right) \times 2}{n_a + n_b}$$

Beispiel: Similar-Text
Zeichenkette a = Beispieltext 1
Zeichenkette b = Biespeil text 2

n_a	n_b	lev(a,b)	max(n_a,n_b) - lev(a,b)	sim	Similar-Text
14	15	6	9	$\frac{9 \times 2}{14+15} = \frac{18}{29}$	0,62

Die berechneten Similaritätswerte beziehen sich immer auf die komplett tokenisierten Segmente, nicht allein nur auf die Suchmaske. Dies führt dazu, dass auch komplett identische Passagen mit einem von 1 abweichenden Wert belegt werden können, wenn sie innerhalb eines größeren Segments benutzt werden. Im nachfolgenden Beispiel ergeben sich die Abweichungen durch den Einschub quick brown.

Beispiel: Similar-Text
Zeichenkette a = The quick brown fox jumps over the lazy dog
Zeichenkette b = The fox jumps over the lazy dog

n_a	n_b	lev(a,b)	max(n_a,n_b) - lev(a,b)	sim	Similar-Text
43	31	12	31	$\frac{31 \times 2}{43+31} = \frac{62}{74}$	0,84

Similar-Text-Berechnungen sind nur bei kurzen Segmenten, wie der Satz-Tokenisierung in eAQUA, sinnvoll, da die Werte mit der Länge der untersuchten Segmente tendenziell abnehmen.

1)

Zum Beispiel die im Bühnenstück von Shakespeare „KING HENRY the Fourth“ abgekürzten „Speaker“-Segmente „North.“ und „West.“ sind Personenbezeichner, keine Himmelsrichtungen.

2)

Solche Listen können sowohl domänenübergreifend, beispielsweise typisch für eine Sprache, als auch domänenspezifisch, beispielsweise typisch für eine Autorenschaft, sein. In eAQUA werden diese Liste anhand von Wortzählungen des Gesamtkorpus erstellt.

3)

In eAQUA ist dies beispielsweise mit der Kookkurrenzanalyse vollzogen worden.

4)

Wenn diese in den Metadaten nicht annotiert sind, ist dies, gerade bei multilingualen Texten, ein nichttriviales Problem, welches häufig durch sprachspezifische (Stich-)Wortlisten gelöst wird.

5)

[Dunning 93]. Dunning, T. „Accurate Methods for the Statistics of Surprise and Coincidenc“. In: Computational Linguistics 19, 1 (1993), 61-74.

6)

Ort / Zeit / Autor: keine Selbstverständlichkeit bei antiken Texten; einige Autorenschaften sind zum Beispiel als **(Pseudo-)** gekennzeichnet oder bei Zeitangaben wird geschätzt, da nur Zeiträume bekannt sind, oder Ortsangaben auf den Sterbeort gelegt

7)

Stoppwortliste: Diese Liste wird für jedes Korpus neu ermittelt, indem alle Wörter gezählt werden.

8)

Die Suchmaske, das Muster, besteht aus nur 5 Termen. Parallelstellen mit doppelt oder mehr Termen ergeben deswegen mehr als eine Suchmaske und eine dementsprechende Anzahl Fundstellen.

9)

[OLIVER 93].Oliver, Ian. Programming Classics: Implementing the World's Best Algorithms. Prentice Hall PTR New York, 1993.

10)

Eine von dem russischen Mathematiker Vladimir I. Levenshtein 1965 eingeführte Methode, zwei Zeichenketten zu vergleichen, indem die minimale Anzahl von Einfüge-, Lösch- und Ersetz-Operationen gezählt wird, um die eine in die andere umzuwandeln.