Hauptseminar Intelligente Systeme
Sprachliche Informationsverarbeitung
Institut für Linguistik, Universität zu Köln
Wintersemester 2006-2007

Ein System zur Textklassifikation

Dozent: Prof. Dr. Jürgen Rolshoven
Referenten: Valentina Rahmanian, Fabian Steeg, Sonja Subicin

Übersicht

- Texklassifikation, allgemein
- Textklassifikation, unser Verfahren

I. Konzept

II. Suffixbäume

III. Systemarchitektur, Implementation

IV. Komponentensysteme

V. Vorführung, Experimente

Defi ition und Abgrenzung

- Definition *Textklassifikation*: Automatische Zuordnung von Texten zu vordefinierten Kategorien
- Ähnlich: *Keyword-Extraction*, hier werden Wörter im Text auf ihre Relevanz für den Inhalt des Textes untersucht (Bsp. Magisterarbeit Alberts), erfordert kein Training, keine Beispiele
- Unterschied zur Textklassifikation: Die Klassen, denen Texte zugeordnet werden, müssen nicht im Text vorhanden sein (ein Artikel über Politik muss nicht das Wort *Politik* enthalten), erfordert Training, z.B. aus Beispielen: Induktives, exemplarbasiertes Lernen

Aufbau von Textklassifikationssystemen

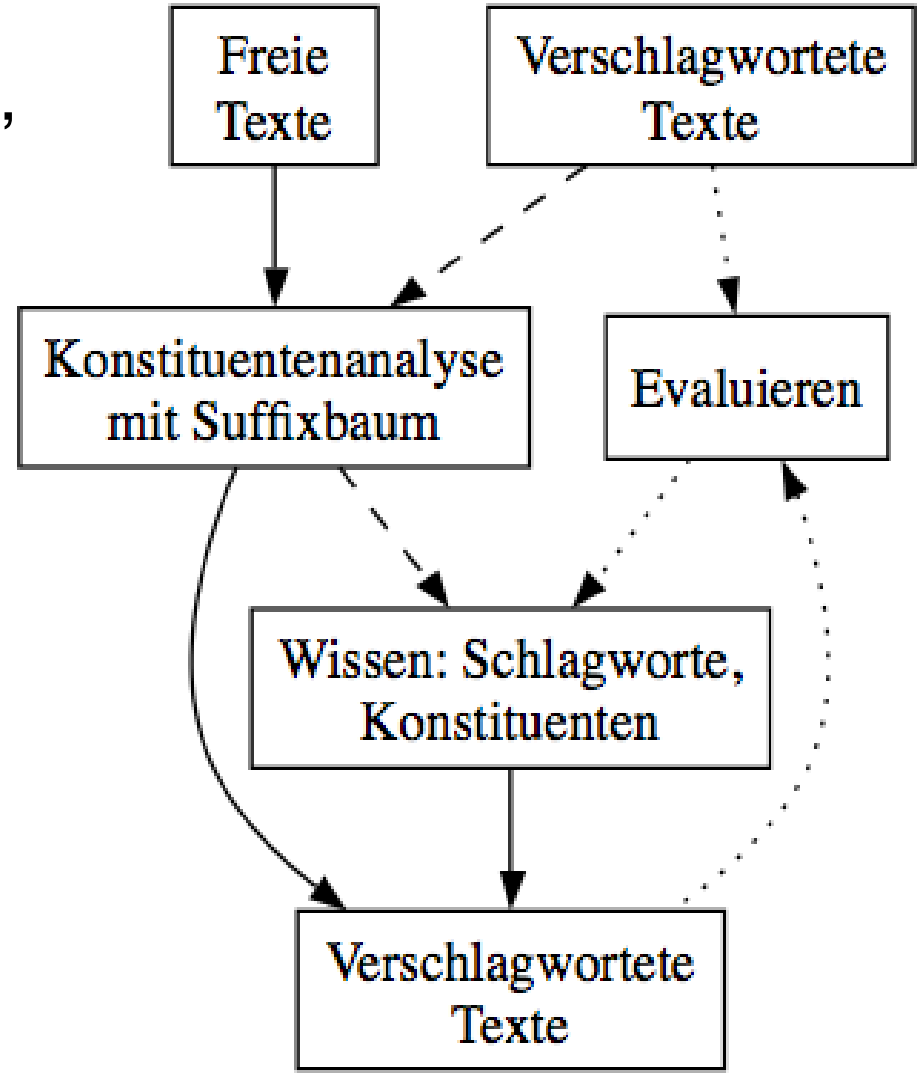
- Zwei Hauptkomponenten (z.B. Brückner 2001): Wissenserwerb (durch maschinelles Lernen oder manuell) und die eigentliche Klassifikation
- Der Wissenserwerb besteht aus Merkmalsberechnung, Merkmalsauswahl und Modellbildung
- TODO Abbildung zur Übersicht
- Ein allgemeines Problem im ML und speziell bei der Textklassifikation ist das *Overfitting:* Nur die Lerndokumente werden richtig klassifiziert

Unser Verfahren

- Grundgedanke: Texte haben das gleiche Thema, wenn sie die gleiche Wörter in den gleichen Kontexten haben (d.h. wenn sie ähnliche Paradigmen enthalten)
- Konzeptuell, was machen wir, Paradigmen
- TODO Beispiel, Abbildung: Paradigem, Texte
- Code, Klassen, UML, JUnit vorführen, praktische Probleme, Hürden etc.

Übersicht

- Als Eingabe dienen schon verschlagwortete, klassifizierte Texte, daraus wird gelernt
- Dann können freie, unklassifizierte Texte klassifiziert werden
- Zur Evaluation werden klassifizierte Texte neu klassifiziert und das Ergebnis mit der ürsprünglichen Klassifikation verglichen

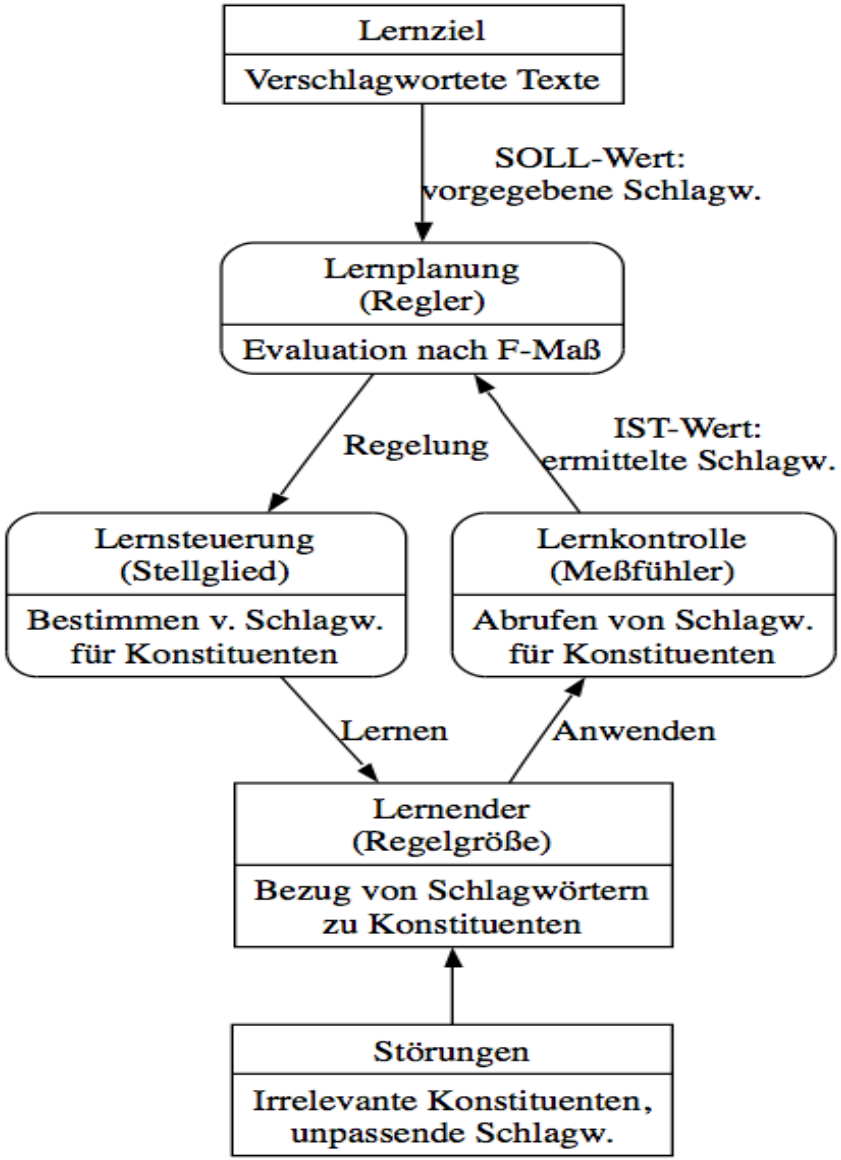


Wissenserwerb bei uns

- Merkmalsberechnung: Erstellung von Paradigmen (durch Suffixbäume, vgl. auch Magisterarbeit Schwiebert)
- Merkmalsauswahl: Die besten Paradigmen (Filtern durch Stopwort-Listen)
- Modellbildung: Bezug von Klassen zu Paradigmen
- Der Wissenserwerb ist eine Form von maschinellem Lernen, Verfahren zum maschinellen Lernen verwenden numerische Repräsentationen der Merkmale, d.h. Hier: Ob und wie sehr ein Paradigma relevant für eine Klasse relevant ist (TODO stimmt das so?)

Maschinelles Lernen

- Felix v. Cube, Kybernetische Grundlgen des Lernens
- Regler: Änderung von Schwellenwerten o.ä. Entsprechend der Evaluation nach F-Maß
- Induktives, exemplarbasiertes Lernsystem (Lernen aus Beispielen, Wörterbuch der Kognitionswissenschaft)



Paradigmen

- Strukturalistische Ideen (syntagmatische Relationen, Ursprung: Harris etc., siehe auch Schwiebert Magisterarbeit)
- Eine strukturalistische Syntaxanalyse
- Paradigmen im Text, daraus andere lernen (wenn gemeinsame Wörter in zwei Paradigmen dann auch die anderen ergänzen), TODO ev. Beispiel, wenn Zeit für sowas

Paradigmen durch Suffixbaeume

- Suffixbaume enthalten in ihrer Struktur Paradigmen
- Suffixbaeume lassen sich effizient erstellen
- Daher koennen mit Suffixbaeumen effizient Paradigmen aus grossen Textmengen erstellen werden (vgl. Schwiebert, SOG)
- Nicht vorhandene Paradigmen die abgeleitet werden koennen

Suffixbäume

- TODO Abbildunug SB für Buchstaben und Wörter, Anwendungsbereich SB traditionell, NLP
- TODO Abbildung oder so für ein Paradigma im Baum
- Paradigmen sind ein sprachunabhängiges Prinzip und die Ermittlung mit Suffixbäumen ebenso

Korpuslinguistik

- Korpuslinguistik: Erstellung und Auswertung von Textkorpora, besondere Bedeutung im NLP, insb. Maschinelles Lernen
- TODO Definition, sampling frame, Ausgewogenheit
- Wenn etwa Nachrichtenartikel klassifiziert werden sollen, sollte das Korpus aus Zeitungsnachrichten bestehen, nicht aus Romanen oder Alltagskonversation
- TODO Web, Crawling, Delicious, Korpora

Delicious

- Delicious: Ein Web-2.0 Tool zum *social bookmarking*
- Ermöglicht Klassifikation von Bookmarks und Bündelung von Klassen
- Solche Bündel könnten bei entsprechenden Inhalten den sampling frames entsprechen, so kann Delicious zum Aufbau von domänenspezifischen Korpora verwendet werden

SALE

- Definition SALE, Sinn, Vorteile, Beispiele
- TODO Abbildung: Crawling 2x, Tokenisierung 2x, Paradigmen (Suffixbaum vs. SOG) 2x, Wortlisten-basiertes Filtern, Lernen (unser Verfahren)
- Sinn: Wiederverwertbarkeit von Komponenten und Ergebnissen: Arbeitsersparnis; Vergleich verschiedener Verfahren (SEMALD)
- z.B. GATE, UIMA, Tesla

Tesla

- Übersicht, Screenshots
- Java EE, Clustering, Standards
- TODO: Wenn Tesla fertig wäre: Komponenten-Diagramm (Abbildung Folie vorher?)

Vorführung

- z.B. Korpus mit Zeitungsartikeln (auf delicious)
- Paradigmen zeigen für vers. Korpora, Filtern, Stopwortlisten

Evaluation (Maße)

- Precision: Wie viele der ermittelten Klassen sind korrekt
- Recall: Wie viele der zu ermittelnden Klassen wurden auch tatsäachlich ermittelt
- F-Maß: Einheitliche Betrachtung von Precision und Recall
- Relative Error: Alternatives Maß zur Evaluation (Goll et al. 2000)

Evaluation
(Was kann man evaluieren)

- Die Qualität eines klassifizierten Dokuments: Vergleich von ermittelten Schlagwörtern und den menschlich vergebenen (Wie viele der zu ermittelten wurden auch ermittelt; Wie viele der ermittelten sind auch relevant)
- Die Qualität das Verfahrens: Mittelwerte für eine Menge klassifizierter Dokumente

Evaluation (Ergebnisse)

- Test mit kleinem Spiegel-Online-Korpus aus 120 Artikeln, Test mit 16 Artikeln, 10.000 Merkmalen, 1 Kategorie pro Artikel
- Ergebnisse (besonders Recall) könnten darauf deuten dass es was bringen könnte
- Ergebnisse bei allgemeineren Korpora mit vielen Kategorien schlechter

Text	Recall	Precision	F-Maß
1	0	0	0
2	1	0,25	0,4
3	0	0	0
4	1	0,25	0,4
5	1	0,5	0,67
6	1	0,5	0,67
7	1	1	1
8	1	0,5	0,67
9	1	0,5	0,67
10	1	0,5	0,67
11	1	0,35	0,5
12	1	0,5	0,67
13	1	0,5	0,67
14	1	0,25	0,4
15	1	0,25	0,4
16	1	1	1
Summe	14	6,85	8,78
Mittel	0,88	0,43	0,55

Mögliche Verbesserungen (sprachunabhängig)

- Qualität der Paradigmen (SOG, Filtern, Mehrwort-Paradigmen, nur mit bestimmten Wörtern, Zusammenfassen)
- Qualität und Quantität der Korpora (z.B. wenige Kat., Zeitungsartikel), HTML-Parsing: Nicht nur Inhalt von Paragraph-Elementen
- Ändern des Algorithmus, Schwellenwert anders, automatisch anpassen. N beste (z.B. 5%) Paradigmen statt festen Wert

Mögliche Verbesserungen (sprachspezifisch)

- Stemming (z.B. WordNet)
- POS-Tagging (z.B. WordNet)
- Semantische Relationen, z.B. Hyperonyme der Wörter im Paradigma (z.B. WordNet, manuell)
- TODO: konkret wie man die nutzen könnte, als Merkmale, zur Verbesserung der Qualität von Paradigmen

Literatur

- Brückner, T.: 2001, 'Textklassifikation', in K. U. Carstensen, C. Ebert, E. Endriss, S. Jekat, R. Klabunde & H. Langer (eds.), *Computerlinguistik und Sprachtechnologie*, Spektrum, Heidelberg, Berlin, pp. 442–447.

- Goller, C., J. Löning, T. Will & W. Wolff: 2000, 'Automatic document classification: A thorough evaluation of various methods', *7. Internationales Symposium für Informationswissenschaft*.

- Yang, Y. & J. O. Pedersen: 1997, 'A comparative study on feature selection in text categorization', in D. H. Fisher (ed.), *Proceedings of ICML-97*, 14th International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, US, Nashville, US, pp. 412–420.

- Hulth, A.: 2003, 'Improved automatic keyword extraction given more linguistic knowledge', *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 03)*, 216–223.

- Cunningham, H. & K. Bontcheva: 2006, 'Computational Language Systems, Architectures', in K. Brown, A. H. Anderson, L. Bauer, M. Berns, G. Hirst & J. Miller (eds.), *The Encyclopedia of Language and Linguistics*, second edn., Elsevier, München.

- McEnery, T.: 2003, 'Corpus Linguistics', in R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford Handbooks in Linguistics, Oxford University Press, Oxford, pp. 448–463.

- Gusfield, Dan: 1997, *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology,* Cambridge University Press.