Prüfungsprotokoll
Diplom NF
Computerlinguistik

Angelika Kimmig
16.06.2004

- Prüfer: Prof. U. Hahn
- Vorlesungen **Natürlichsprachliche Systeme 1 & 2**
- Vertiefungsgebiet **Information Retrieval, Information Extraction, Text Mining** (aufbauend auf dem Hauptseminar "Text Mining" im WS 02/03, Literaturliste s. hinten)

Anmerkung: Durch die Festlegung auf ein seminargebundenes Vertiefungsgebiet unterscheidet sich die Prüfung inhaltlich wahrscheinlich ziemlich stark von nachfolgenden... Grundprinzip: Prüfung als "Fachgespräch" (das ist durchaus wörtlich zu nehmen!), d.h. allgemeiner, überblickartiger Einstieg, gefolgt von Blöcken mit vielen kurzen, aufeinander aufbauenden/vertiefenden/nachhakenden Fragen zu den einzelnen Themenschwerpunkten (in meinem Fall nur da zur Vorlesung, wo diese sich mit der Vertiefungsrichtung überschneidet, d.h. IR-System SMART). Details wurden in vernünftigem Maß gefragt, i.a. genügte ein Beispiel (z.B. ein Ähnlichkeitsmaß definieren, nicht fünf). Wichtiger waren Überblick, Vergleiche, Erklärung grundlegender Prinzipien.

Zum Einstieg: Geben Sie grob an, was Dokumentenretrieval, Informationsextraktion und Text Mining sind und gehen Sie dabei auch schon auf Unterschiede ein.
Dokumentenretrieval/Information Retrieval: sucht relevante Texte zu einer gegebenen Anfrage in einer Dokumentensammlung, keine über die Wortebene hinausgehenden linguistischen Methoden, Benutzer muß selbst Informationen aus Dokumenten holen. IE: Templatestruktur (Wie sieht sowas aus?) als Anfrage, geht in Texte hinein, um Slots zu füllen, verwendet ling. Mittel. TM: weniger konkret gefasste Anfrage, gesucht wird "Neues", Kombination verschiedenster Methoden.
Beschreiben Sie - allgemein oder anhand eines konkreten Systems - wie ein IR-System vorgeht, d.h. gegeben ein Dokument, was passiert?

(konkretes System: SMART) Sämtliche Schritte durchgehen - Wortgrenzen bestimmen (nur kurz angerissen), Stopwörter entfernen (Was sind Stopwörter? Warum brauchen wir die nicht?), morphologische Reduktion (wie? Endungen abtrennen. Beispiel im Englischen? -s, -ing), Termvorkommen zählen, Dokumentenvektoren erstellen (einheitlicher Vektorraum, in dem alle Dokumente der Kollektion dargestellt werden, d.h. jeder Term hat seine feste Position im Vektor)

Was passiert, wenn ich eine Anfrage an das System stelle?

Erzeugung eines Anfragevektors nach gleichem Prinzip, Vergleich des Anfragevektors mit jedem Dokumentenvektor (im naiven Fall) mittels eines Ähnlichkeitsmaßes (Beispiel für ein solches? Cosinus. Formel aufschreiben. Was macht man mit den Ergebnissen? Absteigend sortieren: Ranking)

Welche Möglichkeit hat SMART, wenn man für die ersten 20 Dokumente angibt, welche einem zusagen und welche völlig unbrauchbar sind?

Relevance Feedback - grobes Prinzip: Vektoren der relevanten zur Anfrage addieren, irrelevante abziehen, um Fragevektor in bessere Richtung zu lenken.

Wie bewertet man solche Systeme?

Precision (relevant&gefunden/gefunden) und Recall (relevant&gefunden/relevant). Probleme? Relevanzurteile für gesamte Kollektion. Lösung z.B. fürs Web? Stichprobe/hochrechnen.

Wie geht ein IE-System vor?

Text in Wörter teilen (Probleme dabei?), Extraktion von Einheiten wie z.B. Namen, Geldbeträgen mit regulären Ausdrücken (Bsp. aufschreiben), syntaktische Strukturen (Phrasen, Bäume) aufbauen, in semantische Struktur des Templates übersetzen (wie? Beispielregel? Templatemerging: Probleme? Koreferenzauflösung: Heuristik?)

Was passiert beim Text Mining?

Versuch, neue Informationen, Zusammenhänge etc zu finden in thematisch eingeschränkter Dokumentenkollektion (woher? IR-System)

Was sind Concept Mining und Relation Mining?

Gegeben eine Relation, welche Konzepte erfüllen sie bzw gegeben zwei Konzepte, welche Relationen.

Literatur zur Diplompruefung Informatik (NF Linguistische Informatik)

Grundlagen:

Manning, Christopher D.; Schütze, Hinrich [1999]
Foundations of Statistical Natural Language Processing.
Cambridge/MA, London/England: MIT Press, 1999, 680pp.

Jurafsky, Daniel; Martin, James A. [2000]
Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.
Upper Saddle River/NJ: Prentice Hall, 2000, 934pp.

Information Retrieval, Informationsextraktion, Text Mining

Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier; (Eds.) [1999]
Modern Information Retrieval.
Reading/MA: Addison-Wesley & Longman, 1999

Salton, Gerard; McGill, Michael J. [1983]
Introduction to Modern Information Retrieval.
New York: McGraw-Hill, 1983

Ferber, Reginald [2003]
Information Retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web.
Heidelberg: dpunkt.verlag, 2003, 390pp.

Jackson, Peter; Moulinier, Isabelle [2002]
Natural Language Processing for Online Applications. Text Retrieval, Extraction and Categorization.
Amsterdam, Philadelphia: John Benjamins, 2002, 225pp. (Natural Language Processing, 5).

Pazienza, Maria Teresa; (Eds.) [1997]
Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology.
Berlin: Springer, 1997, 213pp. (Lecture Notes in Computer Science, 1299)

Pazienza, Maria Teresa; (Eds.) [1999]
Information Extraction. Towards, Scalable, Adaptable Systems.
Berlin: Springer, 1999, 165pp. (LNAI Tutorial, 1714)

Gaizauskas, Robert; Wilks, Yorick [1998]
Information extraction: beyond document retrieval.
In: Journal of Documentation, 54, 1998 (1), pp.70-105.

Cowie, J.; Wilks, Y. [2000]
Information extraction.
In: R. Dale, H. Moisl and H. Somers (eds.) Handbook of Natural Language Processing. New York: Marcel Dekker, 2000, pp.241-260.

Merkl, Dieter [2000]
Text Data Mining.
In: R. Dale, H. Moisl and H. Somers (eds.) Handbook of Natural Language Processing. New York: Marcel Dekker, 2000

Appelt, Douglas E. [1999]
Introduction to Information Extraction.
In: AI Communications, 12, 1999 (3), pp.161-172.

Trybula, Walter [1999]
Text mining.
In: Martha E. Williams (Ed.), Annual Review of Information Science and Technology (ARIST), Vol.34: 1999. Medford, NJ: Information Today, 1999, pp.385-419.

Hearst, Marti A. [1999]
Untangling text data mining.
In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. University of Maryland, College Park, Maryland, USA, 20-26 June 1999. San Francisco/CA: Morgan Kaufmann, 1999, pp.3-10.