Hauptstudiumsprojekt WiSe 06/07

Crosslingual retrieval in Web

Walther v. Hahn, Cristina Vertan
{vhahn,vertan}@informatik.uni-hamburg.de

Wozu dient ein Projekt ?

- „Projekte im Umfang von 6 SWS dienen der Bearbeitung größerer theoretischer, konstruktiver oder experimenteller Aufgaben und können in zwei aufeinander folgenden Semestern stattfinden. Ein Projekt kann in Zusammenarbeit mit einem Forschungsprojekt des Fachbereichs oder mit einer Einrichtung außerhalb des Fachbereichs veranstaltet werden.
- Projekte werden in Gruppen durchgeführt und ermöglichen das Erlernen von Gruppenarbeit. Die gesamte Projektgruppe arbeitet auf ein gemeinsames Ergebnis hin.“

Exzerpt aus dem Studienführer „Informatik“, 2004/2005

Womit beschäftigt sich unsere Projekt?

- Crosslingual Information Retrieval ist eine sehr aktuelle Thema, besonders im Bezug mit Webrecherche und eLearning
- Das Hauptstudiumsprojekt wird Ressourcen aus dem EU-Projekt LT4eL (Language Technology for eLearning) benutzen.
- Wünschwert sind Ergebnisse die Hinterher ein Vergleich zwischen konzeptbasierte -Verfahren (LT4eL) uns stochastische Verfahren (Hauptstudiumsprojekt) ermöglichen

Was ist Information Retrieval (IR)

- „Information Retrieval (IR) deals with the representation, storage, organisation of, and access to information items. The representation and organisation of the information items should provide the user with easy access to the information in which he is interested“

(Baeza-Yates, Ribeiro-Neto,
Modern Information Retrieval,
Addisson Wesley 1999)

Warum ist IR so wichtig ?

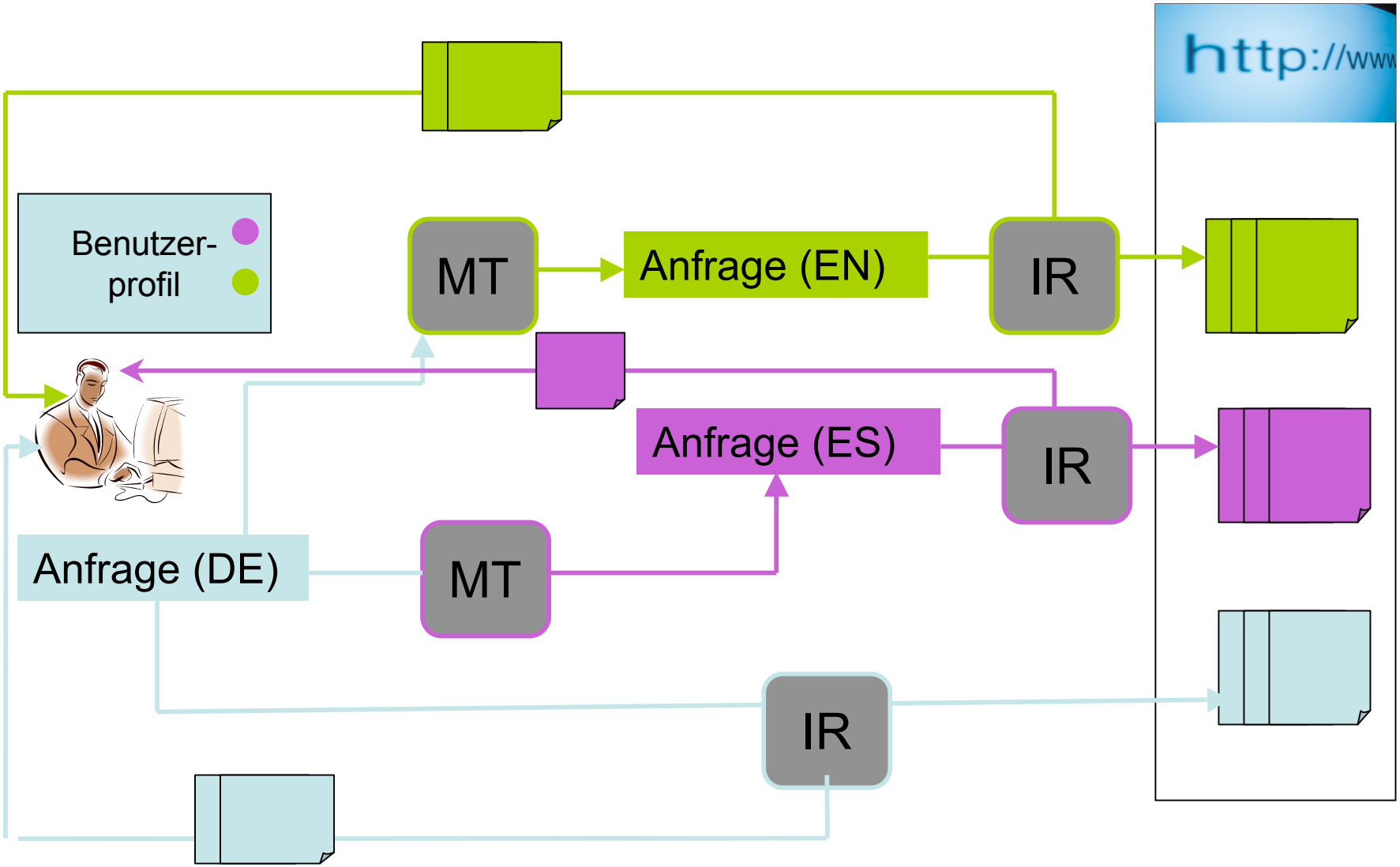
- IR ist Bestanteil vielen Aktionen in unseren alltäglichen Leben:
  - Music Sammlung
  - Organisation von Photos
  - Suche im Web von Literaturquellen für einen Projekt oder Vortrag
  - Informationsorganisation und -suche auf dem eigenen Rechner oder lokalen Netzwerk

Was ist besonders am Web Retrieval ?

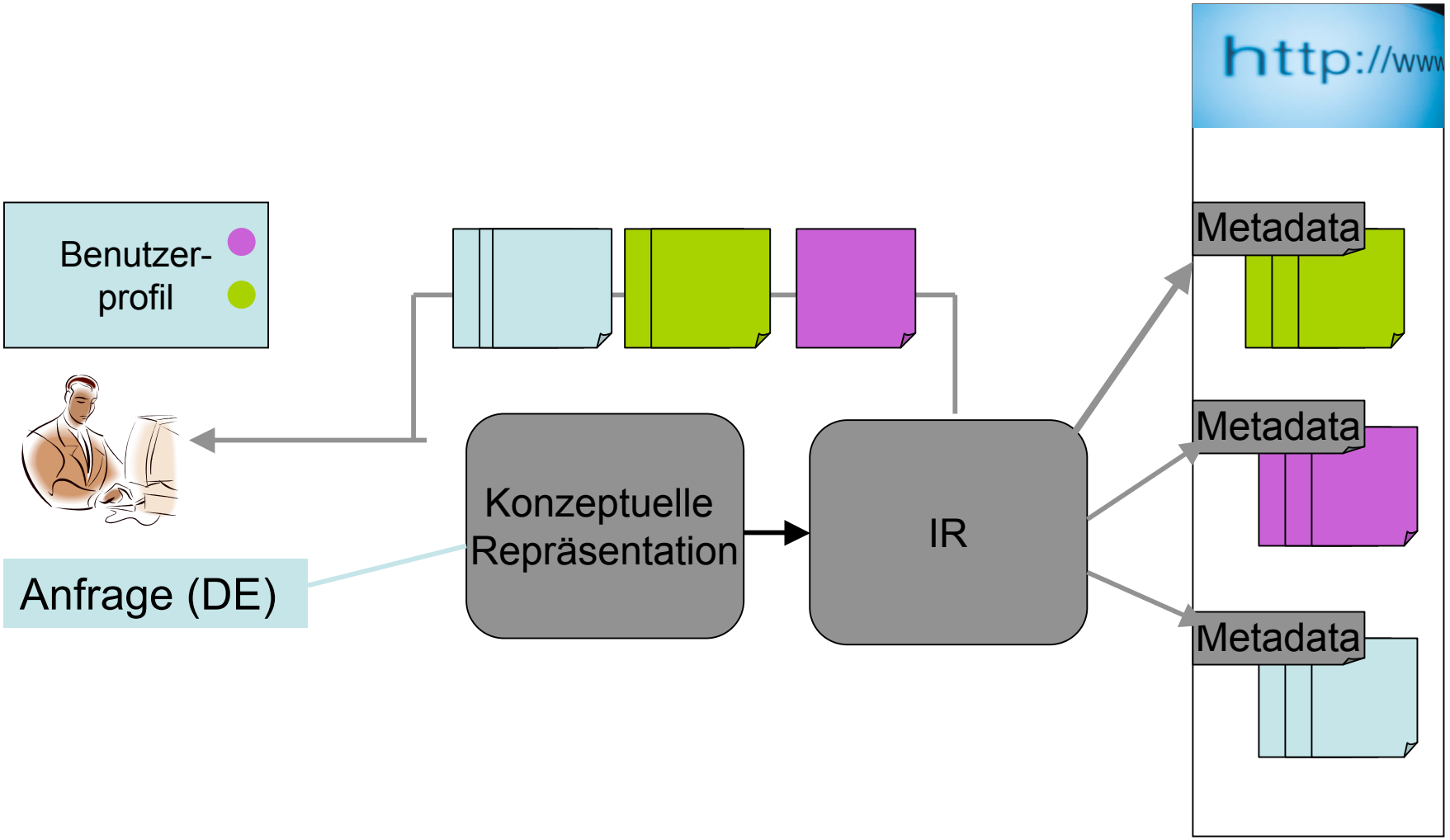
- Man hat wenig Kontrolle über die Datenorganisation
- Dokumenten sind in zahlreichen Formaten vorhanden
- Die Suche variiert von sehr spezifischen bis zu sehr vagen Anfragen
- Viele Benutzer können Dokumente in mehr als eine Sprache lesen, d.h. das Suchverfahren muss die Anfrage auch in anderen Sprachen irgendwie „übersetzen“

Crosslingual Retrieval

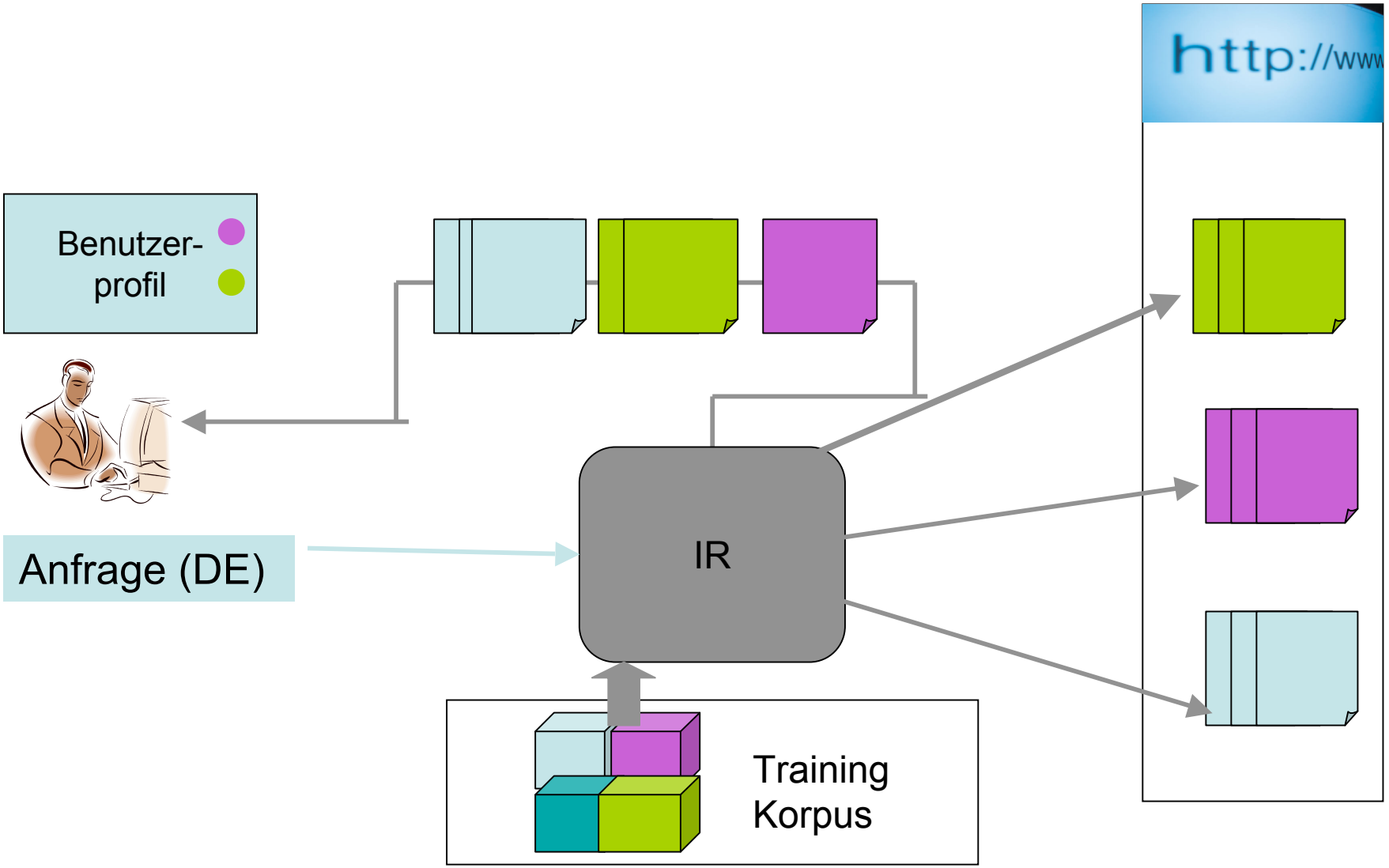
Crosslingual Retrieval -Verfahren -1-
„Übersetzung“



Crosslingual Retrieval -Verfahren -2-
„Semantic Web“



Crosslingual Retrieval -Verfahren -3-
Stochastische Verfahren



Vorgehen

- 25.10 Einführung und Ressourcen-Vorbereitung
- 01.11 Theoretische Grundlagen (IR und MT)
- 08.11 - 20.12 - Implementierung
- 20.12 - Zwischenevaluation
- 10.01 - 31.01 Implementierung
- 31.01 - Vorbereitung End-Präsentation
- 07.02 - End Präsentation

Was wollen wir erreichen ?

- Kenmenlernen von Grunverfahren des Irs
- Evaluation von Ergebnissen:
  - Wieviele Dokumente wurden gefunden ?
  - Wurden die Relevante Dokumente gefunden ?
  - Wie Präzis muss die Übersetzung der Anfrage sein ?
- Gruppen-Arbeit
- Testen und Evaluation des Systems

Scheinkriterien

Teilnahmeschein

- Anwesenheit (maximal 2 Abwesenheiten)
- Implementierung Testen und Evaluation eines Teils des Systems
- Vortrag in der End-Präsentation

Leistungsschein

- Kriterien für Teilnahmeschein +
- Verfassung eines Projektberichts (bis Ende des Semesters)

Ressourcen

- Webseite des Projekts

http://nats-www.informatik.uni-hamburg.de/view/CrossLingIR/WebHome

- Resourcen-Webseite:

http://consilr.info.uaic.ro/uploads_lt4el/

- Login:
  - German: gerLT4eL
  - Dutch: dutLT4eL
  - English: engLT4eL
  - Polish: polLT4eL
  - Portuguese: porLT4eL
  - Czech: czeLT4eL
  - Bulgarian: bulLT4eL
  - Maltese: malLT4eL
- Passwd: elearning

Ressourcen Vorbereitung

- Sammlung von Dokumenten in HTML in mindestens 2 Sprachen (zirka 20 Dokumente / Sprache für die Implementierungsphase und noch 20 für Testen)
- Die Dokumente sollen nicht identisch in allen Sprachen sein (bitte wählen Sie deswegen nicht Calimera Dokumente), aber mit änliche Inhalt
- Analysieren Sie Die Dokumente und überlegen Sie sich zirka 20 Fragen/Patterns die für die Suche benutzt werden können