

Hans Uszkoreit Vorlesung
Informationswissenschaft
und Informationssysteme



Übung zur Vorlesung im Sommersemester 2009

Informationsextraktion & Information Retrieval

**Fachrichtung 4.7 Allgemeine Linguistik
an der Universität des Saarlandes**

Betreuung: Brigitte Jörg

Aufgabe 8: Informationsextraktion (IE)

- 8.1 Was sind Hauptanwendungen von Informationsextraktion?
- 8.2 Nennen Sie weitere Anwendungen in denen IE eine Rolle spielt.
- 8.3 Was ist der aktuelle Stand der IE Technologie? Was kann man heute? Geben Sie mindestens fünf Beispiele.
- 8.4 Was kann man heute noch nicht? Geben Sie mindestens zwei Beispiele.
- 8.5 Mit welchen Methoden arbeiten die gängigen Suchmaschinen (insbesondere Google)? Erläutern Sie kurz deren Stärken und Schwächen.
- 8.6 Was sind die aktuellen Herausforderungen an Suchtechnologien?

Aufgabe 9: Entitätenerkennung / Namensekennung (NER)

- 9.1 Entitätenerkennung ist ein Teilgebiet der Informationsextraktion, bei dem z.B. Eigennamen identifiziert werden. Geben Sie weitere Beispiele von Entitäten.
- 9.2 Auch wenn wir Menschen einen Satz nicht verstehen können, so können wir doch Namen erkennen, die wir nie zuvor gesehen haben. Wir erkennen sie über Muster; der Form oder des Kontexts.

Geben Sie jeweils zwei Beispiele für Muster der Form bzw. für Muster des Kontexts.

Aufgabe 10: Informationsextraktion

10.1 Namensextraktion ist eine Art der Informationsextraktion.

10.2 Welche der Arten können unter Relationsextraktion subsumiert werden? Warum ?

Aufgabe 11: Schritte zur Namenserkennung

Die Namenserkennung wird von mehreren Komponenten getragen. Nennen Sie diese und beschreiben Sie kurz deren Hauptmerkmale.

Aufgabe 12: Relationen

Geben Sie jeweils zwei mögliche Beispiele für

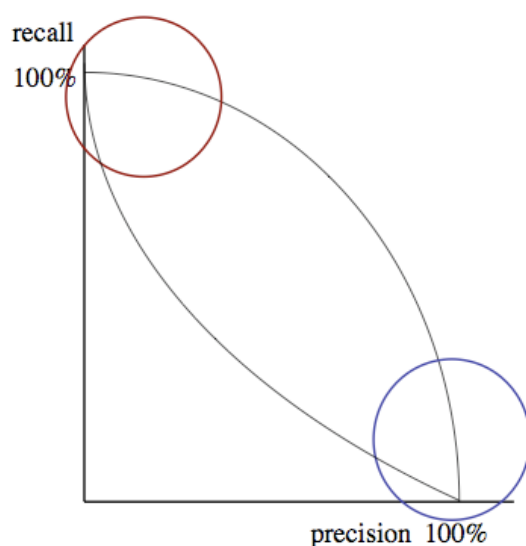
- a) Relationen zwischen Entitäten
- b) Relationen zwischen Ereignis und Datum / Uhrzeit / Ort
- c) Instanzen von Relationen

in Textform und formalisiert.

Aufgabe 13: Information Retrieval

Erklären Sie kurz die Bedeutung von Recall und Precision.

Aufgabe 14: Recall (Trefferquote) & Precision (Genauigkeit)



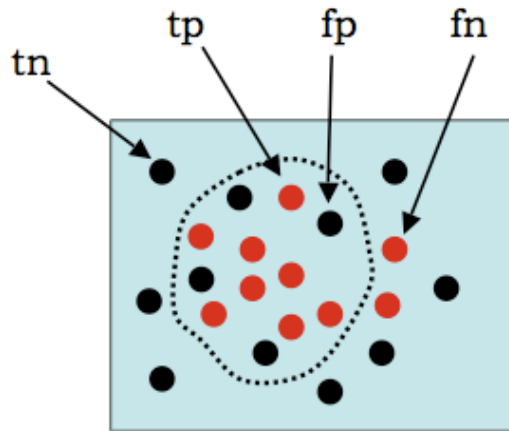
In der Graphik sehen Sie einen rot markierten Bereich sowie einen blau markierten Bereich. Nennen Sie Aufgabenbereiche, die dem jeweiligen Bereich zugeordnet werden können und erklären Sie kurz warum.

Aufgabe 15: Recall & Precision

Formalisieren Sie die nachfolgenden Definitionen in (1) Recall und (2) Precision:

tp = true positives
fp = false positives
tn = true negatives
fn = false negatives

(1) Recall =
(2) Precision =



Aufgabe 16: Recall & Precision

16.1 In einer Sammlung befinden sich 120 Dokumente. Wir wissen dass 20 davon relevant sind. Eine Suche nach relevanten Dokumenten ergibt 25 Treffer, von denen jedoch 7 falsch sind.

Berechnen Sie entsprechend den Recall und die Precision.

16.2 Eine Suche nach Dokumenten ergibt 30 relevante Treffer. Wir können davon ausgehen, dass eigentlich 40 von den 300 Dokumenten in der Datenbank für uns relevant sind.

Berechnen Sie entsprechend den Recall und die Precision.

Aufgabe 17: F-Measure / F-Maß

Erklären Sie kurz die Bedeutung von F-Measure und geben Sie die Formel zu deren Berechnung an. Berechnen Sie anschließend auf Basis der folgenden Werte (Recall=10%; Precision=90%) das F-Maß, bei (1) Gleichgewichtung der beiden Größen sowie (2) bei einer Gewichtung von $\alpha = 0.5, 0.7, 0.9$

Aufgabe 18: **Kompetenzsuche & Analyse**

Nennen Sie (1) wichtige Mittel welche die Kompetenzsuche unterstützen sowie (2) Indikatoren welche mögliche Hinweise auf Kompetenzen beinhalten, und potentielle (3) Quellen die zur Analyse herangezogen werden können. Geben Sie darüberhinaus Beispiele für (4) mögliche Anwendungen einer Kompetenzsuche und -analyse.

Aufgabe 19: **Regelwerke**

Regelwerke werden im Rahmen der Sprachverarbeitung für verschiedenste Aufgaben benötigt. Wie entstehen diese, bzw.; wie werden diese generiert?

Aufgabe 20: **Selbstlernverfahren (Relationsextraktion)**

AM DFKI wurde ein Selbstlernverfahren entwickelt, welches das Erlernen von Relationsextraktionsmustern erweitert. Listen Sie die Hauptschritte dieses Systems für den Bootstrappingprozess und erläutern Sie anhand eines Beispiels dessen Grenzen.

Aufgabe 21: **Small-World & Non-Small-World**

Geben Sie Beispiele für a) Small Worlds und b) non Small Worlds.

Aufgabe 22: **Small-World & Non-Small-World**

Welchen Nutzen haben Kleine Welten für den Bootstrappingprozess? Welche Möglichkeiten gibt es, wenn Kleine Welten nicht verfügbar sind ?

Aufgabe 23: **Bootstrapping -> Recall & Precision**

Ein möglichst hoher Recall ist immer anzustreben, weil selbst bei einer Precision von 1% mit einem 100% Recall noch sehr nützliche Informationen vorhanden sind. Nennen Sie mögliche Hindernisse, die einen hohen Recall bzw. Precision verhindern. Nennen Sie desweiteren Möglichkeiten solche Hindernisse zu überwinden.