

Übungsaufgaben 10

1. Information Extraction

Folgender Absatz stammt aus dem Wall Street Journal (19.02.1993):

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

In diesem Absatz kommen zwei Organisationen vor:

- ORGANISATION-1: New York Times Corporation
- ORGANISATION-2: New York Times Newspaper

Außerdem kommen zwei Personen vor:

- PERSON-1: Russell T. Lewis
- PERSON-2: Lance R. Primis

Ein wirtschaftlich interessantes Template ist "succession" ("Nachfolge"), das folgendermaßen aussieht:

```
<SUCCESSION>  
ORGANIZATION :  
POST          :  
WHO_IS_IN     :  
WHO_IS_OUT    :
```

Das erste Succession-Ereignis in dem Absatz kann folgendermaßen dargestellt werden:

```
<SUCCESSION-1>  
ORGANIZATION : <ORGANIZATION-2>  
POST          : "president"  
WHO_IS_IN     : <PERSON-1>  
WHO_IS_OUT    : <PERSON-2>
```

Dieses Template wurde mit Information aus dem ersten Satz und dem letzten Satz gefüllt. Die Analyserregel besteht aus zwei Teilen, die sich grob folgendermaßen darstellen lassen:

```
[ORGANIZATION] named [WHO_IS_IN] [POST]... [WHO_IS_IN] succeeds  
[WHO_IS_OUT].
```

- a. Insgesamt kommen in diesem Absatz *sechs* Succession-Ereignisse vor. Füllen Sie die Templates für SUCCESSION-2 bis SUCCESSION-6. Verwenden Sie dabei Referenzen auf die ORGANISATIONen und PERSONen wie im Beispiel. Möglicherweise sind nicht alle Slots aller Templates gefüllt.
- b. Geben Sie für jedes Template eine Analyseregeln wie im Beispiel an. Beschränken sich die meisten Analyseregeln auf einen Satz, oder umfassen sie mehrere Sätze. Was heißt das für die Flexibilität der Analyse?

2. Information Retrieval

Information Retrieval funktioniert am besten für Wörter, die nur eine Bedeutung haben, und die nicht zu selten vorkommen. Suchen Sie auf Google nach Dokumentation oder Handbüchern zu den folgenden drei Programmen (auf Deutsch!).

Verwenden Sie jeweils eine Handvoll Suchbegriffe („Handbuch“, „Dokumentation“, „Verwendung“, etc.). Beschreiben Sie, wie gut die Qualität der Ergebnisse war sowie auftretende Probleme, und erklären Sie die Probleme mit Eigenschaften der Programmnamen. Können Sie jeweils die Anfrage erweitern, um die Qualität der Ergebnisse zu verbessern?

- „Preview“ (der OS X PDF-Viewer)
- „Finder“ (der OS X Dateimanager)
- „Graphviz“ (ein Visualisierungsprogramm)
- bnfc (ein Backus-Naur-Form-Konverter)

3. Question Answering

Probieren Sie fünf der Beispiel-Anfragen des QA-Answerbus-Systems aus (<http://answerbus.coli.uni-saarland.de>) und vergleichen Sie die Antworten des Systems mit den Antworten von Google (IR) für dieselben Anfragen.

- a. Hoffentlich sind die Antworten des Answerbus-Systems relevanter... Können Sie Fehler des IR-Systems erkennen, die vom QA-System vermieden werden?
- b. Gibt das QA-System auch falsche Antworten? Spekulieren Sie, welche Probleme zu den falschen Antworten geführt haben könnten. Falls Sie damit Schwierigkeiten haben, könnte Ihnen die Lektüre der FAQ-Seite von Answerbus bzw. des verlinkten EACL 2003-Papiers helfen.