Informationsextraktion

Bestimmte Anwendungen bei der semantischen Verarbeitung erfordern keine tiefe linguistische Analyse mit exakter Disambiguierung (= eine einzige und korrekte Lesart).
Hierzu gehört die Informationsextraktion, die durch folgende Eigenschaften gekennzeichnet werden kann:

- **es können Muster für semantisch interessante Information definiert werden**
- **es muss keine vollständige semantische Repräsentation für Texte erstellt werden.**

Beispiel:
Eine Aufgabe der MUC-5 (1993) bestand in der Heraus-filterung von Information aus Texten über *joint ventures*.

Bridgestone Sports Co. Said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalizes at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and „metal wood“ clubs a month.

Abb. 15.7 zeigt das Ergebnis des FASTUS Systems.
Abb. 15.8 zeigt die kaskadierte Architektur dieses Systems.

TIE-UP-1:	
Relationship:	**TIE-UP**
Entities:	**“Bridgestone Sports Co.“** **“a local concern”** **“a Japanese trading house”**
Joint Venture Company	**“Bridgestone Sports Taiwan Co.”**
Activity	**ACTIVITY-1**
Amount	**NT$20000000**
ACTIVITY-1:	
Company	**“Bridgestone Sports Taiwan Co.”**
Product	**“iron and “metal wood” clubs”**
Start Date	**DURING: January 1990**

Figure 15.7 The templates produced by the FASTUS (Hobbs et al., 1997) information extraction engine given the input text on page 579

Input text:

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and “metal wood” clubs a month.

No.	Step	Description
1	Tokens:	Transfer an input stream of characters into a token sequence.
2	Complex Words:	Recognize multi-word phrases, numbers, and proper names.
3	Basic phrases:	Segment sentences into noun groups, verb groups, and particles.
4	Complex phrases:	Identify complex noun groups and complex verb groups.
5	Semantic Patterns:	Identify semantic entities and events and insert into templates.
6	Merging:	Merge references to the same entity or events from different parts of the text.

Figure 15.8 Levels of processing in FASTUS (Hobbs et al., 1997). Each level extracts a specific type of information which is then passed on to the next higher level.

Company	**Bridgestone Sports Co.**
Verb Group	**said**
Noun Group	**Friday**
Noun Group	**it**
Verb Group	**had set up**
Noun Group	**a joint venture**
Preposition	**in**
Location	**Taiwan**
Preposition	**with**
Noun Group	**a local concern**
Conjunction	**and**
Noun Group	**a Japanese trading house**
Verb Group	**to produce**
Noun Group	**golf clubs**
Verb Group	**to be shipped**
Preposition	**to**
Location	**Japan**

Figure 15.9 The output of Stage 2 of the FASTUS basic-phrase extractor, which uses finite-state rules of the sort described by Appelt and Israel (1997) and shown on page 390.

(1) Relationship: **Entities:**	**TIE-UP** **"Bridgestone Sports Co."** **"a local concern"** **"a Japanese trading house"**
(2) Activity **Product**	**PRODUCTION** **"golf clubs"**
(3) Relationship: **Joint Venture Company** **Amount**	**TIE-UP** **"Bridgestone Sports Taiwan Co."** **NT$20000000**
(4) Activity **Company** **Start Date**	**PRODUCTION** **"Bridgestone Sports Taiwan Co."** **DURING: January 1990**
(5) Activity **Product**	**PRODUCTION** **"iron and "metal wood" clubs"**

Figure 15.10 The five partial templates produced by Stage 5 of the FASTUS system. These templates will be merged by the Stage 6 merging algorithm to produce the final template shown in Figure 15.7 on page 579.

Realisiert sind solche Systeme oft durch endliche Automaten, wie etwa dem zur Erkennung von Organisationen:

Performer-Org	→	**(pre-location) Performer-Noun+ Perf-Org-Suffix**
pre-location	→	**locname | nationality**
locname	→	**city | region**
Perf-Org-Suffix	→	**orchestra, company**
Performer-Noun	→	**Canadian, American, Mexican**
city	→	**San Francisco, London**

Hierdurch können beispielsweise die Organisationsnamen *San Francisco Symphony Orchestra* oder *Canadian Opera Company* erkannt werden.

Es wird keine vollständige syntaktische Analyse durchgeführt, sondern eine domänenabhängige Phrasenerkennung (s. Abb. 15.9).

Die Zuordnung dieser ‚shallow syntax' in entsprechende semantische Muster wird dann durch reguläre Ausdrücke erreicht, wie z.B.:

NG(Company/ies) VG(Set-up) NG(Joint-Venture) with NG(Company/ies)

VG(Produce) NG(Product)

(für den ersten Satz des Beispielstexts)

NG(Company) VG-Passive(Capitalized) at NG(Currency)

NG(Company) VG(Start) NG(Activity) in/on NG(Date)

(für den zweiten Satz des Beispielstexts)

Hierdurch wird das Schema in Abb. 15.10 instanziiert.

Informationsextraktion hat viel mit Information Retrieval zu tun. Insbesondere werden die Bewertungskriterien *Recall*, *Precision* und *F-Measure* übernommen (s. nächste Seite).

METHODOLOGY BOX: EVALUATING INFORMATION RETRIEVAL SYSTEMS

Information retrieval systems are evaluated with respect to the notion of **relevance** — *a judgment by a human* that a document is relevant to a query. A system's ability to retrieve relevant documents is assessed with a **recall** measure, as in Chapter 15.

$$\textbf{Recall} = \frac{\#\text{ of relevant documents returned}}{\text{total }\#\text{ of relevant documents in the collection}}$$

Of course, a system can achieve 100% recall by simply returning all the documents in the collection. A system's accuracy is based on how many of the documents returned for a given query are actually relevant, which can be assessed by a **precision** metric.

$$\textbf{Precision} = \frac{\#\text{ of relevant documents returned}}{\#\text{ of documents returned}}$$

These measures are complicated by the fact that most systems do not make explicit relevance judgments, but rather rank their collection with respect to a query. To deal with this we can specify a set of cutoffs in the output, and measure average precision for the documents ranked above the cutoff. Alternatively, we can specify a set of recall levels and measure average precision at those levels. This latter method gives rise to what are known as precision-recall curves as shown in Figure 17.4. As these curves show, comparing the performance of two systems can be difficult. In this comparison, one system is better at both high and low levels of recall, while the other is better in the middle region. An alternative to these curves are metrics that attempt to combine recall and precision into a single value. The F measure introduced on page 578 is one such measure.

The U.S. government-sponsored TREC (Text REtrieval Conference) evaluations have provided a rigorous testbed for the evaluation of a variety of information retrieval tasks and techniques. Like the MUC evaluations, TREC provides large document sets for both training and testing, along with a uniform scoring system. Training materials consist of sets of documents accompanied by sets of queries (called topics in TREC) and relevance judgments. Voorhees and Harman (1998) provide the details for the most recent meeting. Details of all of the meetings can be found at the TREC page on the National Institute of Standards and Technology Website.

METHODOLOGY BOX: EVALUATING INFORMATION EXTRACTION SYSTEMS

The information extraction paradigm has much in common with the field of information retrieval and has adapted several standard evaluation metrics from information retrieval including **precision**, **recall**, **fallout**, and a combined metric called an **F-measure**.

Recall is a measure of how much relevant information the system has extracted from the text; it is thus a measure of the coverage of the system. Recall is defined as follows:

$$\textbf{Recall:} = \frac{\#\text{ of correct answers given by system}}{\text{total }\#\text{ of possible correct answers in the text}}$$

Precision is a measure of how much of the information that the system returned is actually correct, and is also known as **accuracy**. Precision is defined as follows:

$$\textbf{Precision:} = \frac{\#\text{ of correct answers given by system}}{\#\text{ of answers given by system}}$$

Fallout is a measure of the systems ability to ignore spurious information in the text. It is defined as follows:

$$\textbf{Fallout:} = \frac{\#\text{ of incorrect answers given by system}}{\#\text{ of spurious facts in the text}}$$

Note that recall and precision are antagonistic to one another since a conservative system that strives for perfection in terms of precision will invariably lower its recall score. Similarly, a system that strives for coverage will get more things wrong, thus lowering its precision score. This situation has led to the use of a combined measure called the **F-measure** that balances recall and precision by using a parameter β. The F-measure is defined as follows:

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

When β is one, precision and recall are given equal weight. When β is greater than one, precision is favored, and when β is less than one, recall is favored.