Extraction - Abstraction

- Extraction
  - Extrahiert Teile des Quelltextes, die relevante Information beinhalten. (salience)
  - *Salience* kann auf verschiedenen Ebenen veranschlagt werden
  - Alles was in der ZF vorkommt, ist Teil des Inputs
- Abstraction
  - Die ZF enthält Textteile, die nicht direkt im Quelltext erscheinen
  - Höherer Kompressionsgrad
  - Höherer Abstraktionsgrad
  - Background Konzepte mit Wissenbasis: Thesaurus oder Corpus

Abstraction

1. Analyse und Repräsentation auf semantischer Ebene
2. Diskurs-Repräsentation oder Wissenskonzepte als Grundlage für selection, aggregation und generalisation operations
3. Die neue Repräsentation in NL umformen

4 Methoden für Abstraction

1. Templates
2. Term Rewriting
3. Event relations
4. Concept hierarchies

1. ABSTRACTION FROM TEMPLATES

- Template:
  - Datenstruktur, mit slots und values
- slots:
  - spezifizieren den Typ der Information
  - Benötigen Hintergrundswissen
  - Repräsentieren relevante Information und zeigen, wovon der Text handelt
  - Kompression durch Einschränkung, deckt nur einige Aspekte des Inputs ab

FRUMP (DeJong 1982)

- Ausgangsüberlegung
- Situationen, die in gewisser Weise stereotype, erwartete Ereignisse beinhalten

Sketchy script

Enthält nur die Ereignisse, die in einer gewissen Situation erwartet werden, dass sie eintreffen.

Beispiel eines sketchy scripts

- The demonstratiors arrive at the demonstration location
- The demonstrators march
- Police arrive on the scene
- The demonstrators communicate with the target of the demonstration
- The demonstrators attack the target of the demonstration
- The demonstrators attack the police
- The police attack the demonstrators
- The police arrest the domonstrators

Passendes Script suchen

Sketchy scripts  Artikel

1. Explicit reference

Jedes script hat eine Gruppe von möglichen Wort-bedeutungen. Jedes Wort im Artikel, das diesen Wortsinn erfüllt wird aktiviert

Bsp.

Political demonstration	-> Artikel 1
Demonstration (Experiment)	-> Artikel 2

2. Implicit Reference

Ein *script* kann über ein anderes *script* ausgelöst werden

Bsp:

- ✓ **arrest script** ausgelöst über **crime script**
  Bankraub-Artikel (**crime script**) der im Text *arrest* nicht erwähnt nur von *catching suspects* spricht.
- o Da ein *arrest* gewöhnlich einem *crime* folgt, wird ein *arrest script* ausgelöst.
- o Benötigt Wissen, dass einem Verbrechen eine Gefangennahme folgt, um das **arrest script** auszulösen.

3. Event induced activation

- Schlüsselereignisse *(keyevent)* im script

Bsp: Festnahme ist ein *keyevent* im arrest script

- Wenn ein oder mehrere Schlüsselereignisse des scripts im Input vorkommen, wird das script ausgelöst.

Repräsentation des Scripts in CD

- Die Ereignisse des Scripts werden auf einer semantischen Ebene repräsentiert
- -> Conceptual Dependency (CD)
  Strukturierte Datenrepräsentation mit *slots* und *values* für Scripts, Ereignisse, Wort- und Satzsemantik

Beispiel einer Conceptual Dependency

((<==> (*ATRANS)
MANNER (*FORCED*)
ACTOR (*POLITY*)
OBJECT (*CONT*)
TYPE (*ECONOMIC*)
PART (*SPEC-INDUSTRIE*)
TO (*POLITY*)
FROM (*POLITY*)))

Role-filler pairs

- Die Füllwerte müssen Fälle von spezifischen Konzepten sein
- Der Text wird nach geeigneten Ereignissen durchsucht und die Ergebnisse im Script *instantiated*.
- Bsp: Wortsinn „take“ würde den Agens und das Objekt des Ereignisses suchen und einfüllen.

FRUMP Probleme

- Salience = Information, die bei einem Ereignis erwartet wird. Vorgabe: **Sketchy script.**

Salient Information hängt also ganz von dem ab, was im Script codiert wurde. Wichtige Information, die nicht codiert ist, kann auch nicht gefunden werden.

FRUMP Nachteile

- Neue oder unerwartete Ereignisse?
- Kann nur Inhalte zusammenfassen, welche im Script vorcodiert sind
- CD hat keine systematische Beziehung zu der syntaktischen Struktur eines Satzes und seine symbolisierte Ausgangslage ist sehr generell

FRUMP Innovation

- ✓ Hoher Kompressionsgrad
- ✓ Geht über die Extraktion hinaus
- ✓ Corpus-basierte Methode, die automatisch Templates ausfüllen kann
- ✓ Mit linguistischem Wissen, kann ein ausgefülltes Script in verschiedenen Sprachen generiert werden.

Noch ein Beispiel

- Templates füllen mit einer Kombination von *pattern matching* und *statistischer Analyse*
- Textgenerierung und graphisches Display für die Repräsentation einer Zusammenfassung
- Domain spezifisch

Abstraction für techn. Artikel im Agrikulturbereich (Paice und Jones 1993)

- SPECIES
- CULTIVAR
- PEST
- AGENT
- INFLUENCE
- LOCALITY
- TIME
- SOIL

Pattern machting

- Die Füllwerte der Spalten werden durch Varianten von Mustern (pattern) gefunden
- Bsp. für ein pattern match:
  PEST is a skip-upto-n-words pest of SPECIES
  „A.lolii is a common pest of ryegrass“

Mehrere concept patterns

- *„The effect of mildew seed treatment and foliar sprays used alone or in combination in ‚early' and ‚late' sown Golden Promise spring barley, Aberdeen, 1976 to 1982"*
  - Effect of skip-upto-n-words in SPECIES
  - Effect of skip-upto-n-word in sown SPECIES

Gewichtung

- Unterscheidung zwischen *candidate string* und *filler string* mittels Gewichtung
- -> bester *filler string* für die Textgenerierung

Text mit „canned text“ templates generieren

- **Canned text**

  This paper studies the effect the pest PEST has on the PROPERTY of SPECIES.An experiment in TIME at LOCALITY was undertaken.

- **Zusammenfassung**

  „This paper studies the effect the pest G. pallida has on the yield of potato. An experiment in 1985 and 1986 at York, Lincoln and Peterbourgh, England was untertaken“

2. ABSTRACTION BY TERM REWRITING

- ZF verstanden als linguistischer Prozess, der Symbole und Strings überschreibt.

Macro propositions (Kintsch und van Dijk 1978/79)

1. Jeden Satz des Dokumentes in logischen Ausdrücken, der logische terms enthält, repräsentieren -> **atomic proposition**

- Bsp.: *The rich man lost all hist property,*
- „There is a man“, „he is the same as some previously mentioned man“, „he is rich“, „he lost something“, „the thing he lost was property“, „the property was his“.

Macro proposition

2. Anwendung von Macro-Regeln. Eine Gruppe von **atomic propositions** mit einer einfacheren Repräsentation mit weniger Information ersetzen -> **macro proposition**

- Deletion
- Generalisation
- Construction

Deletion

Peter saw a blue ball

- Atomic proposition:

Peter saw a ball. The ball was blue.

- Macro Proposition:

-> Peter saw a ball

Generalisation

- Peter saw a hawk

-> Peter saw a bird

- Peter saw a hawk, Peter saw a vulture

-> Peter saw birds

Construction

- Peter laid foundations, built walls, built a roof...

-> Peter built a house

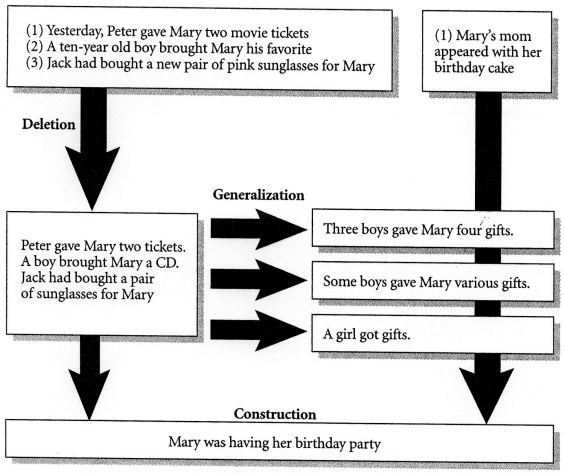

Figure 6.7: Illustration of macro-rules

Vor- und Nachteile

✓ Möglichkeit der Generalisierung

- Benötigt Werkzeuge, die eine Repräsentation auf semantischer Ebene erzeugt
- Generalisierung benötigt eine grosse Wissensbasis und benötigt Einschränkungen

3. ABSTRACTION USING EVENT RELATION

- Semantische und temporale Beziehungen zwischen Ereignissen in einem Text

-> Repräsentation, die *concept coherent* ist (Alterman 1985)

NEXUS (Alterman 1985)

- Semantische Repräsentation eines Textes innerhalb eines event connectivity graphs.

Bsp. The prince spied the pig carrying her laundry to the stream.

The prince was the agent of a spying event whose theme was a carrying event. The agent of the carrying event was a pig, the destination was the stream, and the object was laundry possessed by the pig.

Event Connectivity Graph für:

The prince spied the pig carrying her laundry to the stream.

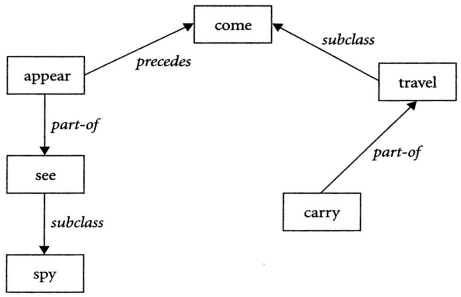

Figure 6.8: Graph for (6).

NEXUS Graph als Ausgangslage für SSS

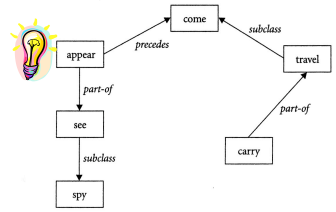

Figure 6.8: Graph for (6).

1. Konzeptuelle Wurzeln der Geschichte finden
   -> Auflisten der Ereignisse
1. Salience entspricht den Knoten, zu dem am wenigsten Kanten führen.
2. Entfernen der Ereignisse in der Liste, die weniger als eine Durchschnittsrelevanz haben.
   -> Restl. Ereignisse sind Basis für ZF

Vor- und Nachteile

- ✓ Generalisierung und Hintergrundswissen kann für die ZF eingeführt werden
- ✓ Abstrakte Konzepte der Wissenbasis liegen der Informationskohärenz zu Grunde
- ✓ Topologie des Graphs wird genutzt um salience zu bestimmen
- ▪ Gebunden an spezifische Domains, dessen Ereignisstruktur bekannt ist.
- ▪ Benötigt eine Wissensbasis, die Ereignisse und ihrer Beziehungen charakterisiert

4. ABSTRACTION USING CONCEPT HIERARCHIES

- Generalisierung basiert auf Hierarchie
- Wissensbasis: Fachbereichkonzepte
- Graphische Repräsentation dessen, wovon der Text handelt

TOPIC (Reimer und Hahn 1988)

1. Parsing von noun phrases,
2. Stüzt sich auf ein Lexikon, welches einer Wissenbasis von Domainkonzepten angeglichen ist
3. Aktivierung der Gewichte von frames, slots und slot values, wenn diese im Text vorkommen -> **concept counting** ergibt *salience*

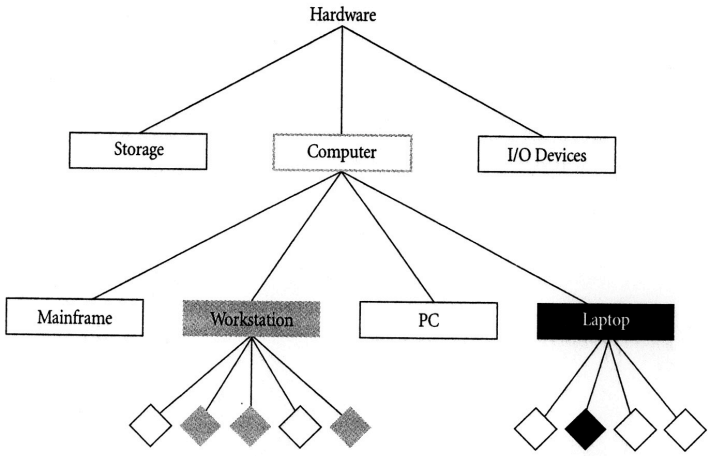

Figure 6.9: Salience based on generalization (from Hahn and Mani 2000)

Vor- und Nachteile

- ✓ Generalisierung kann kontrolliert werden
- ✓ Background Konzepte stammen von einer Hierarchie
- ✓ Bereist existierende Thesauri können verwendet werden
- ■ Wissensbasis für jeden Bereich muss zuerst erstellt werden
- ■ Das Resultat muss für den Anwender lesbar sein

1. Template

- Erfasst Info, die vom Template erwartet werden
- Background Konzepte stammen von Template slots
- Target template wird durch die Domain oder das Genre bestimmt
- Salience basiert auf dem Einfüllen von Werten in slots

2. Term-Rewriting

- Auswählen, Hinzufügen und Zusammensetzen von logischen Ausdrücken auf einer semantischen Repräsentationsebene
- Salience basiert auf concept counting

3. Event-relations

- Salience basiert auf Anzahl von ausgelösten Ereigniskonzepten
- Graph-Zusammenhang wird über die semantische Beziehung definiert
- Background Konzepte notwendig, um die Ereignisse zu verbinden

4. Concept Hierarchie

- Salience basiert auf concept counting
- Generalisierung basiert auf Hierarchie

SYNTHESE

1. Schön drucken
2. Graphisches Output
3. Extraktion

NL Generation für Synthese

1. Bessere Lesbarkeit als Tabellen
2. Kompaktere ZF als Graphik
3. Textplanungs- und lexikalische Strategien für Textkohärenz
4. Ausdrücke der natürlichen Sprache müssen dafür gewählt und kombiniert werden. ->aggregation und generalization während Synthese

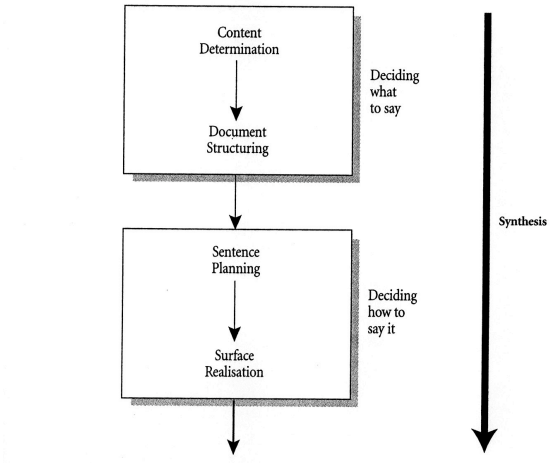

Figure 6.10: Architecture of a generation module

Conclusion

- Generation bietet weitere Möglichkeiten für ZF: compaction, aggragation, generalization
- Generation von strukturierten Daten, getrennt vom Zusammenfassungsprozess nicht günstig
- Corpus: Gewichtung von salience während dem Auswahlverfahren
- Abhängig von Fortschritten in NLG und der corpus-based statistical generation
- Kriterien dafür, was unter einer effektiven Zusammenfassung verstanden wird