Two-Level, Many-Paths Generation (1995)

Kevin Knight und Vasileios Hatzivassiloglou

und

Generation that Exploits Corpus-Based Statistical Knowledge (1998)

Irene Langkilde und Kevin Knight

Vortrag von Johannes Braunias
Universität des Saarlandes, 3. Dezember 2010
im Proseminar „Generierung“
bei Alexander Koller

Generierung

- Für:
  - Wettervorhersage
  - Mensch-Maschinen-Dialog
  - Maschinelle Übersetzung
  - Erklärende Systeme
  - Zusammenfassen
- Quellen:
  - Datenbanken (Wetter)
  - Symbolische Repräsentationen (semantisches Modell)

Beispiel

Semantisches Modell (in AMR*):

```
(m3 / |eat,take in|
      :agent (m4 / |dog<canid|
            :quant plural)
      :patient (m5 / |os,bone|)
            :quant plural))
```

Text:

The dogs eat the bones.

* AMR = Abstract Meaning Representation

Beispiel

Japanischer Text
犬は骨を食べた。

Semantisches Modell:

```
(m3 / |eat,take in|
      :agent (m4 / |dog<canid|
            :quant plural)
      :patient (m5 / |os,bone|)
            :quant plural))
```

Text:

The dogs eat the bones.

Semantisches Modell:

```
(m3 / |eat,take in|
      :agent (m4 / |dog<canid|
            ?  )
      :patient (m5 / |os,bone|)
            :quant plural))
```

Text:

The **dog(s?)** eat the bones.

Lücken in der Eingabe

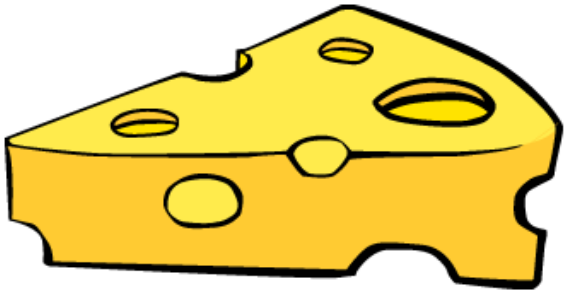
→ Problem!

Schwierigkeiten

- Unterspezifiziertheit in Quelle
- Unzureichende Grammatik
- Lexical Choice-Problem
  - **im** Februar / **am** Freitag / **um** fünf
  - **wie Katz und Maus** / **wie Katze und Feldmaus**
  - You may be obliged **to eat** a poulet. /
    The consumption of poulet by you may be the requirement. /
    It might be the requirement that the chicken **is eaten** by you.

Lücken im Generator
→ Problem!

„Dog oder Dogs?“ – Das ist die Frage.



→ Lücken füllen durch:

- **Default-Werte & Templates –** z.B. immer Singular
  - Regeln manuell erstellen
  - nicht robust
- **Zufällige Auswahl** aus allen Möglichkeiten
  - holprige Ausdrücke – „An earth circles a sun.“

→ geht es besser?

Der neue Weg

Aus unklarer AMR alle Möglichkeiten generieren:

The dogs eat the bone.

The dogs eats the bone.

The dog eat the bone.

The dog eats the bone.

→ Zufall?

Der neue Weg

Aus unklarer AMR alle Möglichkeiten generieren:

The dogs eat the bone. 173
The dogs eats the bone. 154
The dog eat the bone. 153
The dog eats the bone. 194

→ ~~Zufall?~~ **Vergleich der Bigramme mit statistischem Modell!**

Satz 1:		Satz 2:		Satz 3:		Satz 4:	
the dogs …	245 mal	the dogs …	245 mal	the dog …	210 mal	the dog …	210 mal
dogs eat …	230 mal	dogs eats …	2 mal	dog eat …	3 mal	dog eats …	195 mal
eat the …	233 mal	eats the …	201 mal	eat the …	233 mal	eats the …	201 mal
the bone …	169 mal	the bone …	169 mal	the bone …	169 mal	the bone …	169 mal
173		154		153		**194**	

Gemittelte Werte

Die Zahlen und Beispiele sind willkürlich gewählt.

N-Gram-Modelle

„The dog eats all the bones.“

Bigramme (N = 2):

The dog
dog eats
eats all
all the
the bones

Trigramme (N = 3):

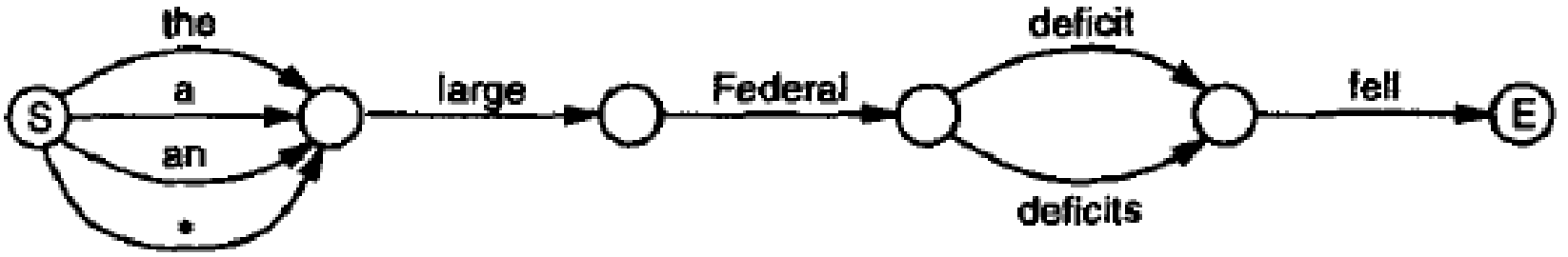
The dog eats
dog eats all
eats all the
all the bones

Erstellen eines N-Gram-Modells:

1. Korpus wird aufgeteilt in N-Gramme
2. Häufigkeiten der N-Gramme im Korpus wird ermittelt

Erzeugen eines Satzes

- Satz mit 7 binären und 2 triären Ambiguitäten:
  - Generator erzeugt $2^7 * 3^3 = 3.456$ Versionen eines Satzes
    → Braucht Platz und Rechenzeit
  - Geht es effizienter?
    → Verbände (lattices):
    Alle Versionen in einer Struktur
    Bsp. für 8 Versionen:



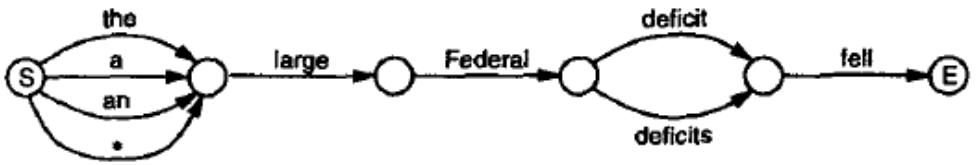
Verbände

Statt Sätzen:

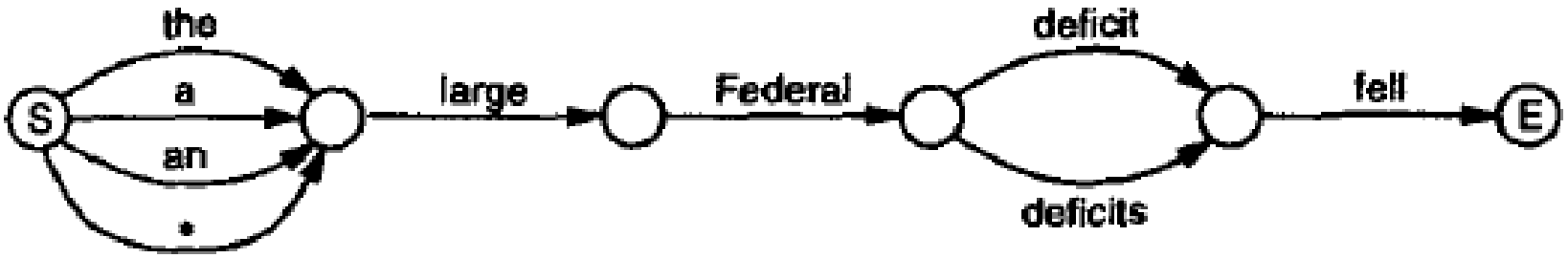
The large Federal deficit fell.
A large Federal deficit fell.
An large Federal deficit fell.
Large Federal deficit fell.
The large Federal deficits fell.
A large Federal deficits fell.
An large Federal deficits fell.
Large Federal deficits fell.

… Verbände:

```
(S (seq
        (or (wrd the) (wrd a) (wrd an) (wrd *))
        (wrd large)
        (wrd Federal)
        (or (wrd deficit) (wrd deficits))
        (wrd fell)
))
```



Verbände



Darstellung als Code:

```
(S (seq
      (or (wrd the) (wrd a) (wrd an) (wrd *))
      (wrd large)
      (wrd Federal)
      (or (wrd deficit) (wrd deficits))
      (wrd fell)
))
```

Wie werden Verbände erstellt?

Semantisches Modell:

(m3 / |eat,take in|
:agent (m4 / |dog<canid|
:patient (m5 / |os,bone|))

Regel:

((x1 :agent) (x2 :patient) (x3 :rest)
→ (s (seq (x1 np) (x3 v-tensed) (x2 np))))

Ablauf:

- Startpunkt ist das Modell.
- Die erste passende Regel wird auf das Modell angewendet.
- Passende Elemente kommen in die Variablen
- Nicht passende kommen in :rest –
  Auf sie werden wieder Regeln angewendet.
- Der Statische Post-Prozessor wählt wahrscheinlichsten Pfad aus

Verband (Bsp. „Titelzeile“):

s (seq (wrd dog) (wrd eats) (wrd bone))
OR
s (seq (wrd dog) (wrd eats) (wrd bones))
OR
s (seq (wrd dogs) (wrd eat) (wrd bone))

Many Paths; Level 1

Statistischer Post-Prozessor

Level 2

Ergebnis:

„Dogs eat bones“

Instanzregel

- Instanz-Regel generiert die Basis-Wörter für die „Blätter“ – die Konzepte
  - Regel **wrd**
  - Attribute :polarity, :quant, :tense und :modal generieren Nomen, Verben und Sonstiges, evtl. verneint, in Zahl, Zeit und Modalitäten

Semantisches Modell:

```
(A / |have the quality of being|
      :DOMAIN (P / |procure|
            :agent (A2 / |American|
                  :quant sg)
            :patient (G / |car|)
      :RANGE (E / |easy, effortless))
```

Wie werden Verbände erstellt?

Semantisches Modell:

```
(A / |have the quality of being|
      :DOMAIN (P / |procure|
            :agent (A2 / |American|)
            :patient (G / |car|)
      :RANGE (E / |easy, effortless))
```

E-Structure:

A2 = (<syn, **lat**>, <syn, lat> …)

= (<NP, (NP (or (wrd **Americans**)
(wrd **American**)
(NP1 {***the Americans***})
))>, …)

Regeln für höhere Ebenen der AMR

- Bilden semantische und syntaktische Rollen auf Verbände ab
  - Rollen sind **agent**, **patient**, … **oblique1**, **oblique2**, … **subject**, **object**, **mod** …

Semantisches Modell:

```
(A / |have the quality of being|
      :DOMAIN (P / |procure|
            :agent (A2 / |American|
                  :quant sg)
            :patient (G / |car|)
      :RANGE (E / |easy, effortless))
```

Rolle1 Rolle2 → Wort1 Wort2

Wie werden Verbände erstellt?

Semantisches Modell:

```
(A / |have the quality of being|
      :DOMAIN (P / |procure|
            :agent (A2 / |American|)
            :patient (G / |car|)
      :RANGE (E / |easy, effortless))
```

Elemente eines Verbandes:

(seq x y …)
(or x y …)
(wrd "word")

E-Structure:

A2 = (<syn, **lat**>, <syn, lat> …)
= (<NP, (NP (or (wrd **Americans**)
(wrd **American**)
(NP1 {***the Americans***})
))>, …)

(NP (seq (wrd **the**) (wrd **Americans**))

Beispiel

```
(m7 / |eat,take in|
    :time present
    :agent (d / |dog,canid|
              :quant plural)
    :patient (b / |os,bone|
                :quant sing))
```

AMR

```
((x1 :agent)
 (x2 :patient)
 (x3 :rest)
 ->
 (s  (seq (x1 np nom-pro) (x3 v-tensed)
          (x2 np acc-pro)))
 (s  (seq (x2 np nom-pro) (x3 v-passive)
          (wrd "by") (x1 np acc-pro)))
 (np (seq (x3 np acc-pro nom-pro) (wrd "of")
     (x2 np acc-pro) (wrd "by") (x1 np acc-pro)))
 (s-ger (seq ...))
 (inf (seq ...)))
```

Regel → …

→ **Verbände**

```
(S (or (seq (or (wrd "the") (wrd "*empty*"))
            (wrd "dog") (wrd "+plural")
            (wrd "may") (wrd "eat")
            (or (wrd "the") (wrd "a")
                (wrd "an") (wrd "*empty*"))
            (wrd "bone"))
       (seq (or (wrd "the") (wrd "a")
                (wrd "an") (wrd "*empty*"))
            (wrd "bone") (wrd "may") (wrd "be")
            (or (wrd "being") (wrd "*empty*"))
            (wrd "eat") (wrd "+pastp") (wrd "by")
            (or (wrd "the") (wrd "*empty*"))
            (wrd "dog") (wrd "+plural"))))
(NP (seq (or (wrd "the") (wrd "a")
             (wrd "an") (wrd "*empty*"))
         (wrd "possibility") (wrd "of")
         (or (wrd "the") (wrd "a")
             (wrd "an") (wrd "*empty*"))
         (wrd "consumption") (wrd "of")
         (or (wrd "the") (wrd "a")
             (wrd "an") (wrd "*empty*"))
         (wrd "bone") (wrd "by")
         (or (wrd "the") (wrd "*empty*"))
         (wrd "dog") (wrd "+plural"))))
(S-GER ...)
(INF ...)
```

Vergleich

```
INPUT

  (H / |have the quality of being|
    :DOMAIN (H2 / |have the quality of being|
              :DOMAIN (E / |eat, take in|
                        :AGENT YOU
                        :PATIENT (P / |poulet|))
              :RANGE (O / |obligatory|))
    :RANGE (P2 / |possible, potential|))

LATTICE CREATED

  260 nodes, 703 arcs, 10,734,304 paths;
  48 distinct unigrams, 345 distinct bigrams.
```

Vergleich

RANDOM EXTRACTION

You may be obliged to eat that there was
the poulet.

An consumptions of poulet by you may be
the requirements.

It might be the requirement that the chicken
are eaten by you.

DEFAULT EXTRACTION

That the consumption of the chicken by you
is obligatory is possible.

STATISTICAL BIGRAM EXTRACTION

1 You may have to eat chicken.
2 You might have to eat chicken.
3 You may be required to eat chicken.
4 You might be required to eat chicken.
5 You may be obliged to eat chicken.

Zwischenbewertung: +

- Keine komplexe Grammatik und Morphologie notwendig: Das *Sprachmodell* bewertet
- Lexikalische Constraints werden indirekt berücksichtigt
- Domain- bzw. Korpus-spezifisches Training liefert gute Ausdrücke für eine Domain

Zwischenbewertung: –

- Fehlendes Agreement
- Statistische Komponente kennt keinen Part of Speech:
  The company planned [the] increase in production.
- Bevorzugt kurze Sätze
- Keine verschachtelten Strukturen möglich:
  John <u>looked</u> it <u>up</u>.

Nitrogen

- NLG-System, basiert auf in JAPANGLOSS verwendetem Generator
- Erweiterte lexikalische und morpholog. Datenbank
- Flexible Eingabe möglich,
  ermöglicht Transformation (Recasting) von AMRs

Lexikalische Datenbank

sell wurde mehrdeutig analysiert:

```
(m6 / (*OR* |sell<cozen| |betray,fail| |cheat on| ))
```

Neue *Ranking*-Spalte in lexikal. DB nutzen:

```
(<word> <part-of-speech> <rank> <concept>)
```

"betray"	VERB	1	|deceive|
"betray"	**VERB**	**2**	**|sell<cozen|**

"sell"	VERB	1	|advertize|
"sell"	VERB	2	|bargain|
⋮	⋮	⋮	⋮
"sell"	**VERB**	**6**	**|sell<cozen|**

- disambiguierte Wörter für ambige Konzepte
- schlechte Auswahl durch Ranking-Info verhindern

Morphologische Datenbank

```
("-child" "children")
("-person" "people" "persons")
("-a" "as" "ae")                ;formulas/formulae
("-x" "xen" "xes")              ; boxes/oxen
("-man" "mans" "men")   ; humans/footmen
("-Co" "os" "oes")
```

- Übergenerierung + Reduktion
- depart → department / departure

Recasting

```
(m8 / |obligatory<necessary|
   :domain (m9 / |eat,take in|
               :agent (m10 / |dog,canid|)))
```

AMR

```
((x1 :rest)
 (x2 :domain)
 ->
 (? (x1 (:new (/ |have the quality of being|)
    (:domain x2) (:range x1)) ?))
```

Recasting-Regel

```
(m11 / |have the quality of being|
   :domain (m12 / |eat,take in|
               :agent (d / |dog,canid|))
   :range (m13 / |obligatory<necessary|))
```

AMR

Recasting

- Verwendet, wenn Eigenschaften bei Instanz-Regel nicht definiert sind (quant, pol, tense, mod)
- Alternative Realisierung
- Ändern der syntaktischen Struktur
  (z.B. einbetten)

Nitrogen – Bewertung

+ Verbesserte Transformation von AMRs in Verbände
  - Erweitertes Lexikon und Morphologie
  - Recasting

- Schwächen
  - *planned [the] trip*
  - Long-distance-Abhängigkeiten schwer behandelbar
  - Keine Generierung von Fragen und Pronomen

Zusammenfassung

- Linguistisch einfache Eingaben
  Übergenerierung in Form von Verbänden
  Selektion mit Hilfe statistischem Sprachmodell aus Korpus
- Gute Ergebnisse bei minimalem Aufwand:
  Knowledge-basiert: komplex und domänenbeschränkt
  Rein statistisch: schlechte Ergebnisse → Hybrid
- Domänen-Wechsel durch Wählen eines anderen Sprachmodells
- robust und skalierbar
- Indirekte Wordsense-Disambiguation

Danke!