

Matematik för språkteknologer

Avslutning

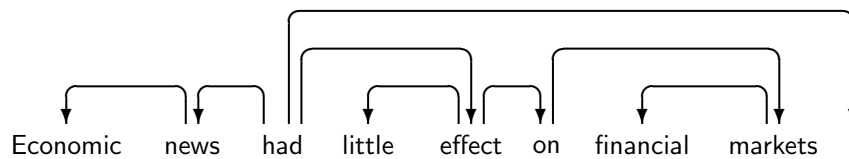
Joakim Nivre

Översikt

- ▶ Matematiska modeller i språkteknologi:
 - ▶ Grafbaserad dependensparsning
- ▶ Kursvärdering:
 - ▶ Muntlig diskussion
 - ▶ Webformulär

Grafbaserad dependensparsning

- ▶ Syntaktisk analys av naturligt språk
- ▶ Syntaktiska representationer baserade på dependensrelationer
- ▶ Representationer formaliserade som riktade grafer



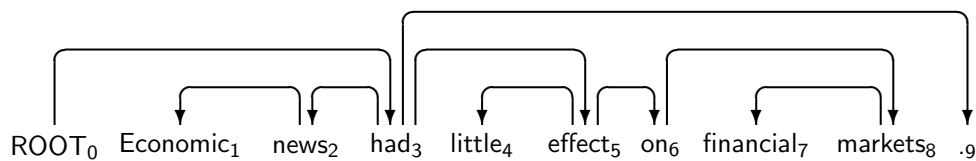
Dependensgrafer

- ▶ En mening S är en sekvens av ord:
 - ▶ $S = w_0, w_1, \dots, w_n$
 - ▶ Ord $w_0 = \text{ROOT}$ är ett fiktivt ord
- ▶ En dependensgraf för S är en riktad graf $G = (V, E)$, där:
 - ▶ $V = \{w_0, w_1, \dots, w_n\}$ (en nod för varje ord, inklusive ROOT),
 - ▶ $E \subseteq V \times V$.

Exempel

$V = \{\text{ROOT}_0, \text{Economic}_1, \text{news}_2, \text{had}_3, \text{little}_4, \text{effect}_5, \text{on}_6, \text{financial}_7, \text{markets}_8, \text{.9}\}$

$E = \{(\text{ROOT}_0, \text{had}_3), (\text{news}_2, \text{Economic}_1), (\text{had}_3, \text{news}_2), (\text{had}_3, \text{.9}), (\text{had}_3, \text{effect}_5), (\text{effect}_5, \text{little}_4), (\text{effect}_5, \text{on}_6), (\text{on}_6, \text{markets}_8), (\text{markets}_8, \text{financial}_7)\}$



Villkor på dependensgrafer

- ▶ En dependensgraf G antas normalt uppfylla följande villkor:
 - ▶ G är (svagt) sammanhängande (dvs. motsvarande oriktade graf är sammanhängande).
 - ▶ G innehåller inga cykler.
 - ▶ Ingen nod i G har en ingrad högre än 1.
 - ▶ Noden $w_0 = \text{ROOT}$ är en rot (har ingrad 0).
- ▶ Av detta följer att en välformad dependensgraf är ett riktad träd med en unik rot $w_0 = \text{ROOT}$.

Dependensparsning

- ▶ Problem:
 - ▶ Input: Mening $S = w_0, w_1, \dots, w_n$ ($w_0 = \text{ROOT}$).
 - ▶ Output: Dependensgraf $G = (V, E)$ för S .
- ▶ Metod:
 - ▶ $\text{Parse}(S) = \text{argmax}_G P(G|S)$
- ▶ Delproblem:
 - ▶ Inläring:
 - ▶ Hur skatta sannolikheten $P(G|S)$ för godtyckliga G och S ?
 - ▶ Inferens
 - ▶ Hur beräkna $\text{argmax}_G P(G|S)$ givet en sådan skattning?

Sannolikhetsmodell

- ▶ Antagande 1:
 - ▶ De ingående bågarna i en dependensgraf är (statistiskt) oberoende av varandra.
 - ▶ Då följer att: $P(G|S) = \prod_{(w_i, w_j) \in E} P((w_i, w_j)|S)$.
- ▶ Antagande 2:
 - ▶ En båge (w_i, w_j) är (statistiskt) oberoende av alla ord utom w_i och w_j .
 - ▶ Då följer att: $P(w_i, w_j|S) = P((w_i, w_j)|w_i, w_j \in S)$.
 - ▶ Och därmed: $P(G|S) = \prod_{(w_i, w_j) \in E} P((w_i, w_j)|w_i, w_j \in S)$.

Skattning av modellparametrar

- ▶ Modellen kräver att vi skattar sannolikheten

$$P((w_i, w_j) | w_i, w_j \in S)$$

för alla ord w_i, w_j i godtyckliga meningar S .

- ▶ Detta kan vi göra genom att räkna den relativa frekvensen av (w_i, w_j) i meningar S som innehåller w_i och w_j :

- ▶ $P((w_i, w_j) | w_i, w_j \in S) = \frac{C((w_i, w_j))}{C(w_i, w_j \in S)}$

- ▶ Observera:

- ▶ Denna typ av skattning förutsätter en syntaktisk analyserad korpus (trädbank).
 - ▶ I praktiken måste vi ta hänsyn till att vissa ord förekommer mycket sällan (eller inte alls) även i en stor korpus.

Sammanfattning

- ▶ Syntaktisk struktur kan representeras av riktade grafer
- ▶ Syntaktisk analys kan realiseras med en sannolikhetsmodell
- ▶ Modellens parametrar kan skattas med hjälp av korpusstatistik