

Tekniker för storskalig parsning

Introduktion

Joakim Nivre

Uppsala Universitet
Institutionen för lingvistik och filologi
`joakim.nivre@lingfil.uu.se`

Kursöversikt

- ▶ Kursnamn: Tekniker för storskalig parsning
- ▶ Kurskod: 5LN450
- ▶ Undervisning:
 - ▶ Tider: Vecka 36–43, i huvudsak måndag och onsdag 13–15
 - ▶ Lokal: Turing (9-2042)
- ▶ Lärare: Joakim Nivre (joakim.nivre@lingfil.uu.se)
- ▶ Kurshemsida: <http://stp.lingfil.uu.se/~nivre/5LN450/>

Kursplan – lärandemål

- ▶ Kursen tar upp regelbaserade och datadrivna metoder för storskalig automatisk morfologisk och syntaktisk analys av naturliga språk.
- ▶ Efter avslutad kurs skall studenten för att förtjäna betyget Godkänd minst kunna:
 - ▶ redogöra för problem som uppstår vid storskalig grammatisk analys;
 - ▶ redogöra för och i någon mån tillämpa standardmetoder för automatisk morfologisk analys inklusive ordklasstagning samt utvärdera dessa;
 - ▶ redogöra för och i någon mån tillämpa standardmetoder för automatisk syntaxanalys inklusive chunkning, ytsyntaktisk parsning och fullständig parsning samt utvärdera dessa;
 - ▶ redogöra för skillnaden mellan regelbaserade och datadrivna modeller för analys av naturligt språk, samt ge exempel på båda typerna för morfologisk och syntaktisk analys och diskutera för- och nackdelar med respektive typ;
 - ▶ formulera och avgränsa problemställningar inom automatisk morfologisk och syntaktisk analys samt lösa problem och skriftligt och muntligt redogöra för dessa på ett vetenskapligt korrekt sätt.

Kursplan – litteratur

- ▶ Två böcker:
 - ▶ Jurafsky and Martin, *Speech and Language Processing*
 - ▶ Kübler, McDonald, and Nivre, *Dependency Parsing* [PDF]
- ▶ Fyra artiklar:
 - ▶ Nivre, On Statistical Methods in Natural Language Processing [PDF]
 - ▶ Nivre, Two Notions of Parsing [PDF]
 - ▶ Nivre, Two Strategies for Text Parsing [PDF]
 - ▶ Nivre, Statistical Parsing [PDF]
- ▶ Ett bokkapitel:
 - ▶ Roark and Sproat, Computational Approaches to Morphology and Syntax, Kapitel 1 [Kopia]

Kursens uppläggning

- ▶ Teoridel (**bredd**):
 - ▶ Översikt av regelbaserade och datadrivna metoder för morfologisk och syntaktisk parsning
 - ▶ Individuella uppgifter
- ▶ Projektdel (**djup**):
 - ▶ Utveckling av en robust dependensparser för obegränsad svensk text med hjälp av datadrivna metoder
 - ▶ Grupparbete

Teorikursen

- ▶ Föreläsningar:
 - ▶ Grundbegrepp (F1–F3)
 - ▶ Dependensparsning (F4–F5)
 - ▶ Probabilistiska grammatiker och modeller (F7, F9, F11)
 - ▶ Finita automater för morfologisk och syntaktisk analys (F13)
- ▶ Litteratur:
 - ▶ Två böcker [1 PDF]
 - ▶ Fyra artiklar [PDF]
 - ▶ Ett bokkapitel [Kopia]
- ▶ Inlämningsuppgifter:
 - ▶ Frågor som besvaras skriftligt till varje delområde
 - ▶ Särskilda VG-uppgifter
 - ▶ Sista inlämningsdag: 7 november

Projekt delen

- ▶ Laborationer:
 - ▶ Grundläggande optimeringsmetodik och preparering av data
 - ▶ Grundläggande särdrag
 - ▶ Särdragsinteraktion
 - ▶ Utvärdering av korrekthet och effektivitet
- ▶ Resurser:
 - ▶ MaltParser (<http://maltparser.org>)
 - ▶ Swedish Treebank
 - ▶ Diverse hjälpprogram för utvärdering och datapreparering
- ▶ Redovisning:
 - ▶ Muntlig redovisning i vecka 42–43 (R14–R15)
 - ▶ Skriftlig rapport (ca 5 sidor) senast **7 november**
 - ▶ Särskilda VG-uppgifter

Examination

- ▶ Kursen examineras genom
 - ▶ skriftliga inlämningsuppgifter på kurslitteraturen,
 - ▶ ett projektarbete som redovisas muntligt och skriftligt.
- ▶ I båda fallen finns
 - ▶ obligatoriska uppgifter för betyget godkänt (G),
 - ▶ frivilliga uppgifter varav minst två tredjedelar måste fullgöras för betyget väl godkänt (VG).
- ▶ Sista inlämningsdag för alla uppgifter är **7 november**.

Kursvärderingar

- ▶ Kursen har getts en gång tidigare (HT 2009).
- ▶ Överlag positiv utvärdering:
 - ▶ Helhetsintryck: 4,6
 - ▶ Övriga aspekter: 4,2–4,8
- ▶ Specifika synpunkter:
 - ▶ För många inlämningsuppgifter?
 - ▶ Differentiering av projektuppgiften

Storskalig parsning – en introduktion

Vad menar vi med storskalig parsning?

- ▶ Parsning:
 - ▶ Indata: Text (eller tal)
 - ▶ Utdata: Strukturell analys
- ▶ Flera nivåer:
 - ▶ Morfologi (lemmatisering, ordklasstaggning)
 - ▶ Syntax (frasstruktur, dependensstruktur)
- ▶ Olika typer av storskalighet:
 - ▶ Täcka en stor del av språket – obegränsad text
 - ▶ Analysera stora mängder text

Vad är problemet?

- ▶ Antag att vi har ett småskaligt system:
 - ▶ En kontextfri grammatik med ett tjugotal regler
 - ▶ En effektiv parsningsalgoritm, t.ex. CKY eller Earley
 - ▶ Fungerar utmärkt för ett fragment av språket
- ▶ Varför kan vi inte bara lägga till fler regler?
 - ▶ Svårt att täcka hela språket – problem med robusthet
 - ▶ Fler regler ökar flertydigheten – problem med disambiguering
 - ▶ Fler regler ökar komplexiteten – problem med effektivitet

Begrepp och notation

- ▶ Formell grammatik:
 - ▶ G = grammatiken
 - ▶ $L(G)$ = det (sträng)språk som definieras av G
 - ▶ $\text{Parse}(G, x)$ = parseträd för x enligt G
 - ▶ **Observera:** $\text{Parse}(G, x) = \emptyset$ om $x \notin L(G)$
- ▶ Naturligt språk:
 - ▶ L = det naturliga språket (t.ex. svenska)
 - ▶ $\text{Parse}(L, x)$ = parseträd för x i L
 - ▶ **Diskutera:** Hur känner vi $\text{Parse}(L, x)$?
- ▶ Grammatiken approximerar det naturliga språket:
 - ▶ $L(G) \approx L$
 - ▶ $\text{Parse}(G, x) \approx \text{Parse}(L, x)$
 - ▶ **Diskutera:** Hur bra kan approximationen bli? Hur vet vi det?

Robusthet

- ▶ Vad händer när $x \notin L(G)$?
 - ▶ $x \in L$ – bristande täckning
 - ▶ $x \notin L$ – robusthet i snäv mening
- ▶ Svår gränsdragning:
 - ▶ Hon har inte sett honom.
 - ▶ Hon har inte sett han.
 - ▶ Hon inte har sett honom.
 - ▶ Hon inte sett honom.
 - ▶ Hon har inte sett honnom.
 - ▶ Hon har inte setthonom.
- ▶ **Diskutera:** Hur vill vi att en parser ska hantera dessa fall?

Disambiguering

- ▶ Vad händer när $|\text{Parse}(G, x)| > 1$?
 - ▶ $|\text{Parse}(L, x)| > 1$ – genuin ambiguitet
 - ▶ $|\text{Parse}(L, x)| < |\text{Parse}(G, x)|$ – övergenerering
- ▶ Observationer:
 - ▶ Övergenerering ökar drastiskt med grammatikens storlek
 - ▶ Praktiska tillämpningar kräver normalt disambiguering
 - ▶ Genuin ambiguitet är sällsynt i kontext
- ▶ **Diskutera:** Hur vill vi att en parser ska hantera dessa fall?

Effektivitet

- ▶ Parsningstid bestäms av tre faktorer:
 - ▶ Parsningsalgoritmens komplexitet, t.ex. $O(n^3 \cdot |G|)$
 - ▶ Meningens längd = n
 - ▶ Grammatikens storlek = $|G|$
- ▶ Observationer:
 - ▶ Parsningstiden ökar linjärt med grammatikens storlek
 - ▶ Parsningstiden enligt ovan inkluderar inte disambiguering
- ▶ **Diskutera:** Hur lång tid får det ta att parse en mening?

Så vad kan vi göra?

- ▶ Robusthet:
 - ▶ Öka täckningsgraden om möjligt
 - ▶ Lätta på kraven (soft constraints)
- ▶ Disambiguering:
 - ▶ Eliminera omöjliga (eller osannolika) analyser
 - ▶ Rangordna kvarvarande analyser (ofta statistiskt)
- ▶ Effektivitet:
 - ▶ Reducera komplexiteten om möjligt
 - ▶ Använda approximeringsalgoritmer

Och vad tar vi upp på kursen?

- ▶ Grundbegrepp:
 - ▶ Representationer, arkitekturer och utvärdering (F2)
 - ▶ Modeller, algoritmer och tekniker (F3)
- ▶ Dependensparsning – fördjupning i projekt
 - ▶ Transitionsbaserad parsning (F4)
 - ▶ Grafbaserad parsning (F5)
- ▶ Probabilistiska grammatiker och modeller
 - ▶ Probabilistiska kontextfria grammatiker (F7)
 - ▶ Generativa modeller (F9)
 - ▶ Diskriminativa modeller (F11)
- ▶ Finita automater för morfologisk och syntaktisk analys (F13)