GÖTEBORGS UNIVERSITET
SPRÅKDATA
Algoritmisk textanalys
Mats Eeg-Olofsson
Sep -77

Algoritmisk textanalys - en presentation

Inom projektet Algoritmisk textanalys håller vi på med att utarbeta formaliserade metoder för grammatisk analys av autentisk svensk text. Ett av projektets syften är praktiskt – vi vill konstruera ett fungerande programsystem som på ett ekonomiskt sätt kan analysera stora textmassor. Existensen av ett sådant system är väsentlig för verksamheten inom Logoteket, det nationella serviceorgan som tillhandahåller maskinläsbara texter och textbearbetningar. Arbetet ger givetvis också teoretiskt relevanta resultat. Man kan t.ex. peka på Staffan Hellbergs inom projektet utarbetade formella beskrivning av svenskans morfologi. Andra lingvistiskt intressanta regelsystem som ligger till grund för analysen är exempelvis Jerker Järborgs ytstruktursyntax. Analyssystemets utformning har också teoretiskt intresse som uttryck för en perceptionsstrategi.

Konkreta delmål för projektarbetet just nu är dels att disambiguera all homografi i den inmatade texten, dels att förse den med en enkel syntaktisk strukturbeskrivning. En praktiskt betydelsefull biprodukt av systemets verksamhet är alltså en lemmatisering av texten. Den syntaktiska ytstrukturbeskrivningen har givetvis ett värde i sig, men kan också tjäna som utgångspunkt för en djupare syntaktisk-semantisk analys.

I dagens läge är det ju knappast möjligt att uppnå dessa delmål på helt automatisk väg. Orsaken tycks vara att det är svårt att med maskinen efterlikna människans förmåga att använda extralingvistiska data även vid en rent "formell" grammatisk analys. Vi förutser därför att en av komponenterna i systemet blir en mänsklig informant, en lingvist som kan granska maskinens lösningsförslag och eventuellt komma med korrektioner. Vi vill emellertid gärna se hur bra den helt automatiska analysen kan bli innan vi bestämmer hur interaktionen lingvist-maskin skall utformas.

Algoritmisk textanalys - en presentation
Mats Eeg-Olofsson
Proceedings of NODALIDA 1977, pages 27-30

Medan andra projekt, framför allt sådana som syftar till någon form av textförståelse, i stor utsträckning har övervunnit svårigheterna vid den grammatiska analysen genom en långtgående integrering av syntaktisk och semantisk bearbetning av texten, vill vi i Algoritmisk textanalys se hur långt man kan komma med rent formella hjälpmedel. Statistiska data om frekvenser för ord och ordförbindelser, hämtade från projektet Nusvensk Frekvensordbok (NFO), kommer att användas på flera sätt. Syftet är att bygga upp ett slags sannolikhetsmodell, som för varje alternativ analys av en mening ger ett mått på analysens rimlighet.

Strategin för textanalysen blir i korthet följande:

Steg 1: junkturanalys och förberedande morfologisk analys.

Detta steg är helt automatiskt. Det utförs delvis för att spara tid och minnesutrymme för bearbetningarna i de följande stegen. Indata är en opreparerad text. Utdata är samma text uppdelad i meningar och försedd med ordklass- och böjningsangivelser för vissa ord.

Svårigheten vid skiljeteckensanalys är ju framför allt att skilja mellan förkortningspunkt och meningspunkt. För att göra denna analys säkrare används en lista från NFO3 över vanliga meningsinledande fraser. De ord som i detta steg förses med grammatiska uppgifter är (vanliga) heterografer och kvasiheterografer (formellt homografa ord, t.ex. "är", där en av homografkomponenterna är helt dominerande). Även ord som ingår i vissa frekventa "konstruktioner" (grammatiskt välformade rekurrenta ordförbindelser) disambigueras därigenom (nästan) säkert och blir därför goda utgångspunkter för homografsepareringen i de följande stegen.

Steg 2: morfologisk analys och syntaktisk ytstrukturanalys.

I detta steg kommer flera olika processer att samverka på ett sätt som vi ännu inte har utformat mera detaljerat. Utdata från steg 1 underkastas först en morfologisk analys, där de ord som inte redan har en grammatisk märkning slås upp i ett lexikon över stammar. Återstoden av ordkropparna undersöks sedan av en morfologisk grammatik som anger möjliga böjningsändelser och fogar för de ord som

tillhör stammens paradigm. Sammansatta ord uppdelas alltså i sina beståndsdelar. Ofta nog kan en stam ha flera alternativa paradigmnummer. Dessa undersöks (och presenteras) då i en ordning som är baserad på de textuella frekvenserna för motsvarande lemman i NFO2. Flertydighet uppkommer också genom att det är morfologiskt möjligt att segmentera längre sammansatta ord på många olika sätt.

Utdata från den morfologiska analysen är en riktad graf, där de olika bågarna representerar alternativa homografkomponenter i de analyserade orden (jämför Martin Kays "chart"). Syntaxkomponenten väljer nu ut strängar av homografkomponenter och märker de ingående orden med avseende på de konstituenttyper de kan ingå i. Strängen kan sedan (i allmänhet på flera sätt) uppdelas i en följd av kontinuerliga "ytstrukturkonstituenter". Varje sådan ytkonstituentsträng värderas därefter genom att de enheter som finns med i ett lexikon över konstituentfraser tilldelas ett högt värderingstal. Detta är en pseudosemantisk kontroll, eftersom de fraser som är frekventa nog att vara belagda i lexikonet är semantiskt selekterade. Ytkonstituenterna underkastas sedan en intern kontroll och värdering; härvid används bl.a. kongruensfenomen. Efter den interna evalueringen följer en extern. Värderingsgrunder härvidlag är regler för en menings totala sammansättning, för ordningen mellan ytkonstituenter och för antalet konstituenter i en viss position. Dessutom utnyttjas information i ett särskilt subkategoriseringslexikon, som t.ex. för verb skall ange hur många och vilka syntaktiska argument som verbet brukar ta. Slutligen sammanvägs de olika evalueringarna av varje ytkonstituentsträng till ett totalvärde, som uttrycker analysens rimlighet.

Steg 3: relationell syntaktisk analys.

I detta steg, som ännu bara befinner sig på planeringsstadiet, skulle man kunna söka efter relationella kategorier som subjekt, objekt osv. bland ytkonstituenterna.

Det återstår att närmare utforma evalueringsreglerna och fastställa värden på de ingående parametrarna. Detta kommer att kräva mycket experimenterande. Betydelsefull för systemets effektivitet blir också återkopplingen mellan de olika analysprocesserna. När skall man

t.ex. avbryta den syntaktiska analysen av en homografkomponentsträng som verkar dålig och välja en annan homografkomponentsträng i stället?

En alternativ systemlösning som något diskuterats inom projektet utgår från texter lagrade i en form som föreslagits för Logotekets ändamål. Texterna finns då i databaser, där man via länkar har direkt tillgång till såväl hela texten som samtliga belägg på varje grafordstyp, initial- och finalalfabetisk sortering av ordtyperna osv. Denna lagringsform är säkert fördelaktig för en användare som vill studera texten intensivt. Den borde också kunna ge vissa fördelar för den algoritmiska textanalysen. Varje grafordstyp behöver bara analyseras en gång. Man har också möjlighet att arbeta med större kontexter än meningar. Vid bestämning av ordklass och böjning för ord som saknas i lexikon kan det vara värdefullt att förfoga över alla belägg i texten på de okända orden. Emellertid torde en sådan metod vara ganska resurskrävande i tid och minnesutrymme. För närvarande verkar det mera angeläget att utveckla den andra systemlösningen, som använder sekventiell bearbetning mening för mening.

Programmeringen sker i SIMULA för IBM-anläggningen vid Göteborgs Datacentral. För vissa interaktiva bearbetningar av lexika och regelsystem används BASIC vid Språkdatas NOVA-anläggning.

Referenser:

Allén, S. et al.: Nusvensk Frekvensordbok 1-3 (NFO1-3).

Kay, M.: The MIND System (i Rustin (ed.): Natural Language Processing, New York 1973).