Automatisk igenkänning av nominalfraser i löpande text

Björn Rauch
Stockholm

Abstrakt

I uppsatsen redogörs för en samling algoritmer för automatisk nominalfrasmarkering i löpande text. Algoritmerna bygger på varandra och använder den redan utförda analysen. De är därmed enkla och inte tidskrävande. Algoritmerna kan grovt indelas i två grupper: den första gruppen markerar **kärnnominalfraser** eller **minimala nominalfraser**. För svenskan rör det sig i stort sett om bestämningar, som står till vänster om huvudledet (substantivet). Den andra gruppen av algoritmer markerar **maximala nominalfraser**. Här lägger man alltså till prepositionsfraser, infinitivkonstruktioner, m m. Den sista gruppen har inte tagits upp här. Indatan har hämtats ur tidningar, böcker och andra publikationer på svenska. Texterna taggades med ordklasser och morfologiskt markerade, grammatiska kategorier, men för övrigt använder algoritmerna ingen lexikal information. (Även om mindre ordlistor kunde förbättra resultatet avsevärt; t ex en lista över substantiv som bestämmer mängden av någonting och som förekommer i appositioner (i exemplen: *par* och *antal*): *ett par minuter*, *ett antal människor*. Utan semantisk information kan man inte avgöra om det rör sig om en eller två NP.) Indatan innehåller ungefär 12 000 kärnnominalfraser. En vidareutveckling av programmet kan vara att jämföra meningar med liknande struktur (samma finita verb) för att skapa ett valenslexikon (huvudsakligen för verb, men substantiv och adjektiv skulle också kunna vara med).

Allmänna principer

Som man ser i inledningen är uppgiften väldigt komplex. Framför allt om man tänker på att det är en korkad dator som skall utföra arbetet. Därför är det nödvändigt att splittra problemet i små delproblem som lättare kan lösas. Samtidigt kan de olika delarna av programmet ta hänsyn till redan utfört arbete, vilket ytterligare underlättar analysen.

En annan fördel med denna indelning är, att man kan följa principen att inskränka den grammatiska informationen i programmets olika delar, för att se vilka grammatiska kategorier som är nödvändiga respektive onödiga för analysen. Vidare skulle programmet vara tolerant mot 'ogrammatiska' nominalfraser, som:

1 *det stort huset*
2 *den hela frågan*

Tyvärr medför detta också en del problem:

3 *... hörde den blinde statsrådet*

Proceedings of NODALIDA 1993, pages 207-215

Exempel 3 markerar algoritmen förmodligen som ett verb plus en NP. Men på det viset blir det troligen lättare att upptäcka felaktigt markerade NP (som i 3). Däremot skulle det vara besvärligt att fastställa med en strikt algoritm att 1 och 2 är NP:er.

Det är följaktligen omöjligt att använda enbart frasstrukturregler eller rewrite-rules, utan att reglerna blir mjuka. Medan frasstrukturregler får problem med "nästan"-nominalfraser, säger den andra typen av regler: "kanske är det en nominalfras".

Programmet delades upp i två delar, nämligen regelbasen, som innehåller de lingvistiska reglerna, och algoritmen, som utvärderar indatan med hjälp av regelbasen. Detta åtskiljande förenklar granskningen och förbättrar därmed de lingvistiska reglerna.

Definition av minimal nominalfras

Målet att markera nominalfraser på enbart morfologiska grunder ter sig faktiskt som ett olösbart problem. Men om man begränsar sig till den delmängd som jag kallar **kärnnominalfraser** (nuclear nominal phrase NNP), blir uppgiften vettigare. Begreppet kärnnominalfras kan definieras med utgångspunkt i begreppet nominalfras genom inskränkningar av densammma.

I kategorin nominalfras ingår många komplexa konstruktioner, som gör meningar tvetydiga. Visserligen kunde man klara av ett antal av dessa konstruktioner på grund av ordföljden (alltså syntaxen), men i botten ligger väl huvudsakligen semantiska (och pragmatiska) kriterier, som upplöser dessa tvetydigheter. Det är i första hand två konstruktioner som skall uteslutas med detta argument: prepostionsfraser som (efterställd) bestämmning till nominalfrasen och samordnade nominalfraser. Här följer exempel på dessa (hakparenteser används för att markera nominalfrasgränser; ibland indiceras dem för att förtydliga vilka som matchar varandra):

4a *[Fredrik] tog [bussen] till [Odenplan].*
4b *[Fredrik] tog [[bussen] till [Odenplan]].*
5a *[1 [2 [3 Hans bedömning 3] av [3 krisen 3] 2] och [2 moderaternas förslag 2] 1] skulle komma fram i [artikeln].*
5b *[1 [2 Hans bedömning 2] av [2 [3 krisen 3] och [3 moderaternas förslag 3] 2] 1] skulle komma fram i [artikeln].*

A- och b-versionerna är båda möjliga analyser av strukturen i nominalfraser och båda meningar ändrar betydelsen. Det finns å andra sidan meningar, där kontexten utesluter den ena eller andra tolkningen:

6 *[Sverige] slog [Finland] i [ishockey].*
7 *[Per] tittade på [[figuren] av [trä]].*

För konjunktioner är situationen annorlunda. Här är det svårt att avgöra **vad** som samordnas. Ordföljdskontexten kan vara sådan att konjunktionen står mellan två nominalfraser (som i exempel 5 ovan) men det samordnas inte nominalfraserna, utan exempelvis två satser.

Det finns en tredje konstruktion som utelämnas, nämligen relativa bisatser. Huvudanledningen här är att bisatser 'döljer' en mängd nominalfraser, som därmed går förlorade för en test av analysen. Å andra sidan kan det vara svårt att avgöra var den relativa bisatsen slutar. Om man tittar på hur relativa bisatser används, ser man visserligen att de antingen slutar där meningen tar slut eller vid nästa finita verb (som inte är bisatsens finita verb). Andra möjligheter undviks nästan alltid, fast de förekommer. Anledningen till denna preferens borde vara att det även för en människa är svårt att förstå dessa meningar.

Sammanfattningsvis är det alltså just de bestämningar som i sin tur innehåller nominalfraser som utesluts.

Första algoritmen: Kontextoberoende analys av orden

Här gäller det att skapa utgångspunkten för den följande analysen. Det är därför av nytta att "övergenerera", dvs markera hellre för många kärnnominalfraser än för få. Om alla möjliga kärnnominal-fraser markeras, kan man i nästa steg koncentrera sig på att avgöra om "äkta" nominalfraser genererades. Det verkar däremot vara svårare och mer tidskrävande att hitta nominalfraser i en ordsträng.

Analysen använder ingen syntaktisk information utan genomför en rent morfologisk analys, dvs den granskar varje ord för sig, utan att ta hänsyn till kontexten. Det tycks vara ganska hopplöst med tanke på den komplexa uppgiften att excerpera nominalfraser. Men idén är att man kan säga för olika ordklasser var de kan hamna i en kärnnominalfras eller om de överhuvudtaget kan förekomma i kärnnominalfraser. Prepositioner och verb t ex förekommer aldrig i kärnnominalfrasen. Likaså kan man påstå att där det finns en determinerare (artikel, demonstrativa pronomen) finns även en nominalfras. Vidare kan man säga att nominalfrasen möjligen börjar på vänstra sidan (alltså att det möjligen finns en NP-gräns till vänster om determineraren). Substantiv står som huvudled i en nominalfras och i en kärnnominalfras längst till höger. Reglerna anger precis dessa förhållandena. De anger möjliga NNP-gränser runt orden i meningen. Här följer ett exempel (fler belysande exempel ansluter i nästa avsnitt):

8 *Flickan kysste den snälla pojken.*

Nu granskas varje ord och i enlighet med reglerna sätts parenteser (möjliga NP-gränser) ut (anm: Understrykning visar vilket ord, som genererar parenteserna.):

8' [*Flickan*]] *kysste* [[*den* [snälla [*pojken*]] .

Det ser lite tokigt ut för det finns för många och onödiga paranteser. Men varje ord producerar ju parenteser och "bryr sig inte om" vad de andra orden gör. Nu ansluter steget som sammanfattar de möjliga NP-gränserna till "riktiga" NNP-gränser. Här finns det tre regler:

(1) Följer två vänsterparenteser ('[') på varandra och finns det endast ord emellan (inga högerparenteser!), stryk parentesen till höger.
(2) Följer två högerparenteser (']') på varandra och finns det endast ord emellan (inga vänsterparenteser!), stryk parentesen till höger.
(3) Använd reglerna (1) och (2) successivt, tills det inte längre går.

Resultatet av 8' blir efter upprepade användning av (1) – (3):

8'' [*Flickan*] *kysste* [*den snälla pojken*].

Denna mening markerades alltså alldeles rätt. Den är naturligtvis enkel och motsvarar inte alls vanlig text, som tidningsartiklar, skönlitteratur o dyl. I det följande avsnittet skildras vilka resultat man kan uppnå och vilka problem verkliga texter åstadkommer för algoritmen.

Det fanns vissa problem som programmet inte klarade av. Nu följer en klassificering av de vanligaste misstagen:

- **appositioner**

Programmet antar att varje substantiv är ett kärnled och eftersom den enda information som används i princip är ordklassen, blir det svårt att avgöra om något är apposition eller inte. En möjlig lösning är att sammanfatta två substantiv som direkt följer på varandra. Men för det första klaras inte 10 och dessutom uppstår nya svårigheter med dubbelt transitiva verb (13) (Appositioner diskuteras i avsnitt 6 mer ingående). Exempel:

9 *[ett par] [minuter]*
10 *[ett stort antal] [gulliga katter]*
11 *[Sovjetledaren] [Michail] [Gorbatjov]*
12 *[mannen] [Kalle]*
13 *[Fredrik] gav [Kalle] [boken].*

- **partikelverb, satsadverbial, gradadverb**

Här måste man nämna en känslig punkt i utgångsmaterialet. Ordklassen adverb är för odifferentierad. I kategorin ingår ord som är mycket litet kopplade med varandra såväl syntagmatiskt som paradigmatiskt. Bl a hör till klassen partiklar (*på, uppe*) och gradadverb (*mycket, ganska, lite*) och de kan inte urskiljas från adverb som *också, givetvis, inte* osv. Detta leder till lustiga fel:

14 *[Ändå] tog [man] [det] [lugnt].*
15 *Rör [ner kryddor] och [salt].*
16 *Men [man] rör [ju bara ihop] [en deg], ...*
17 *[Ofta] räcker [det] att strö [ut ungefär] [2 msk mjöl] i [ett tunt lager] ...*

Delvis lyckades programmet att avskilja satsadverbial och partiklar (16), vilket är positivt eftersom har man först isolerat problemet ... Men på köpet får man även att gradadverb, som står först i en nominalfras (17), borttagits. Detta händer emellertid inte ofta. Däremot har satsadverbialen och partiklar mer än femtio procent andel i felen som programmet gjorde.

- **efterställda possessiv, attribut osv**

Ibland förekommer det att ett possessivpronomen eller en adjektivfras som vanligen föregår kärnledet i nomnalfrasen följer efter (i exemplen markeras endast det intressanta fallet):

18 *Säg mig , [flicka lilla] , har du en bra man?*
19 *[Pappa min] , när går vi till Grönan?*

- **nominalfraser med komplexa adjektivfraser**

Med detta menas en konstruktion som huvudsakligen förekommer i kanslispråket och är väldigt markerad i svenskan. Det handlar om en adjektivfras vars bestämmning är en prepositionsfras:

20 *Man föreslog därför [ett för hästsporten gemensamt riksspel].*
21 *[en, i jämförelse med en mer homogen hyresmarknad, högre hyresnivå]*

Dessa nominalfraser markeras på följande sätt:

20 *Man föreslog därför [ett] för [hästsporten] [gemensamt riksspel] .*
21 *[en,] i [jämförelse] med [en mer homogen hyresmarknad,] [högre hyresnivå]*

Andra Algoritmen: Parsning av nominalfraser

Själva kärnan består av en parser för nominalfraser och en "anti-parser" som bedömer om en ordsträng inte är en nominalfras. Parsern undersöker endast ordföljden och inte om orden är rätt böjda (t ex: "... *och detta sista Viggenplanet som* ..."). Till anti-parsern används just den kunskap om strukturen av sådana icke-nominalfraser som diskuteras i det föregående avsnittet. Proceduren för en nominalfras går till så här:

Först körs np-parsern. Den finner nominalfrasen korrekt eller säger att den inte kan avgöra det (tolkar det på det viset):

		parserns svar
22	*den snälla pojken*	är en nominalfras
23	*ut ungefär*	vet ej
24	*ner kryddor*	vet ej

22 godkänns som nominalfras och algoritmen slutar, medan exempel 23 och 24 måste vidare analyseras. Det är anti-parsern som nu får avslöja icke-nominalfraser:

		anti-parserns svar
23	*ut ungefär*	är ej nominalfras
24	*ner kryddor*	vet ej

Det blir över sådana nominalfraser som varken parsern eller anti-parsern definitvt kunde peka ut (23). Nu börjar programmet "anpassa" kärnnominalfrasen genom att ta bort ett ord i taget från den vänstra sidan och kollar igen om den nya strängen är en nominalfras, alltså:

24 *ner kryddor* → **24'** *kryddor*

Nu börjar parsningen om igen och då kommer parsern att godta 24' som nominalfras. Resultatet för 22 - 24 kan man alltså sammanfatta så här:

22 *[den snälla pojken]*
23 *ut ungefär*
24 *ner [kryddor]*

Tredje Algoritmen: Appositioner

Detta tredje steg bör betraktas som ett **försök** att hitta appositioner. Som antytts i avsnitt 4 kan man inte urskilja appositioner på enbart morfologiska grunder. Det är t o m så att meningar är tvetydiga och först på semantisk/pragmatisk nivå upplöses ambiguiteten:

25 *Ett är statsministerns medvetna ljugande inför [konstitutionsutskottet] [1985].*
26 *Ett är statsministerns medvetna ljugande inför [konstitutionsutskottet 1985].*

Det tyder på att en algoritm med dessa hårda restriktioner kommer att göra många fel och frågan är om man överhuvudtaget skall ha med ett sådant steg i analysen eller om man inte skulle inskränka begreppet kärnnominalfras mera.

Det visar sig att appositioner har en egenskap utöver att delarna (nominalfraserna) följer på varandra. Den andra nominalfrasen saknar nämligen determinerare och är indefinit eller nominalfrasen är ett egennamn. Det intressanta är att indefinita nominalfraser utan determinerare inte används så ofta och framför allt inte i kontexten direkt efter en annan nominalfras.

Resultatet av denna parser är blandat. Siffrorna (se avsnitt 7 nedan) visar att fler appositioner hittades än nominalfraser som inte är appositioner. Här följer exempel på felmarkeringar (endast ordsträngar som programmet markerade som apposition visas):

27 *Ikväll sluter en majoritet i [Sveriges riksdag Uganda] till sin bröst.*
28 *Men nu sprids i stället [de forna socialdemokratiska sympatisörerna vind] för våg.*

Detta vittnar om att appositioner inte taggas tillfredställande. Därför kommer en statistik med, och en utan, det tredje steget att presenteras i nästa avsnitt.

7. Resultat

K/F (≈ recall) och K/G (≈ precision) anger procentuellt andelen korrekta jämfört med facit respektive andelen korrekta jämfört med de genererade. Om alla siffror är 100% är resultatet perfekt, om det genereras korrekta men för få NP kommer högerledet att vara 100%. Ett perfekt vänsterled betyder att alla NP har genereras plus lite till

(maximal metod). Normalt bör endast antalet NP:er studeras, men antalet höger- och vänstergränser ("[", "]") kan vara intressanta vid utvecklingen av olika metoder.

Här följer siffrorna för kärnnominalfraser. I statistiken ingår 19 texter på vardera drygt 2000 ord av olika slag såsom lagtexter, romaner, tidningsartiklar och andra. (Exemplen i uppsatsen har – bortsett från exemplen i algoritm-beskrivningarna – hämtats ur dessa texter.) Sammanlagt innehöll materialet 42024 ord i 2517 meningar.

TABELL 1: Resultat efter andra parsern.

	Facit (F)	**Genererade (G)**	**därav Korrekta (K)**	**K/F %**	**K/G %**
NNP	12450	13011	11566	92.9	88.9
"["	12450	13011	12129	97.4	93.2
"]"	12450	13011	12261	98.5	94.2
Ord i NP	20048	19775	19653	98.0	99.4

TABELL 2: Resultat efter tredje parsern (appositioner).

	Facit (F)	**Genererade (G)**	**därav Korrekta (K)**	**K/F %**	**K/G %**
NNP	12450	12408	11719	94.1	94.4
"["	12450	12408	12019	96.5	96.9
"]"	12450	12408	12145	97.6	97.9
Ord i NP	20048	19775	19653	98.0	99.4

8. Slutsats

Syftet med detta taggningsprogram är att demonstrera vad som kan åstadkommas med enkla metoder och med strikta principer. Trots detta är andelen korrekta nominalfraser hög. Det finns en svag punkt som måste nämnas. Kritiken riktar sig mot materialet, nämligen att indatan innehöll den rätta "tolkningen" av orden. Exempel:

genom:	preposition, adverb, substantiv
maskar:	verb, substantiv
väg:	verb, substantiv
bara:	adverb, subjunktion, adjektiv
den, det, ...	determinerare, pronomen

osv

För att avgöra ordklassen måste man analysera hela satsen. Men även här finns möjligheten att köra en liknande enkel analys som vi gjorde för nominalfraser. Den har en väldigt hög träffsäkerhet.

Nu är det endast frågan om vad som kommer härnäst? Som det redan antytts ovan betraktas endast den inre strukturen av en nominalfras. Den direkta konsekvensen är att fortsätta utöka informationsmängden som ställs algoritmen till förfogande. Här blir det bara att granska nominalfrasens kontext. Exempel:

- Efterföljs en nominalfras direkt av prepositionen *av* + nominalfras sammanfattar allt till en enda nominalfras.
- Börjar meningen med en nominalfras, lägg till allt som följer tills det finita verbet kommer.

Formaliserat blir reglerna (NP står inte för en bestämd nominalfras):

- NP + *av* + NP → NP
- satsbörjan + NP + flera ord + finit verb → (satsbörjan +) NP + finit verb

Dessa regler kan (i nästan samma form) översättas till datorspråk. Det bör anmärkas att analysen kommer att generera maximala nominalfrasgränser.

Referenser

Ejerhed, Eva, Gunnel Källgren, Ola Wennstedt and Magnus Åström. 1992. *The Linguistic Annotation System of the Stockholm – Umeå Corpus Project.* Department of General Linguistics, University of Umeå.

Källgren, Gunnel. 1984. *Automatisk excerpering av substantiv ur löpande text. Ett möjligt hjälpmedel vid datoriserad indexering?* Institutet för Rättsinformatik, Juridiska Institutionen, Stockholms universitet.

Källgren, Gunnel. 1992. *Making maximal use of surface criteria in large-scale parsing: the MorP-Parser.* Institutionen för Lingvistik, Stockholms universitet.

Thorell, Olof. 1977. *Svensk Grammatik.* 2:a upplagan, Norstedts, Stockholm.