En homografseparator baserad på sannolikhet

Gunnar Eriksson
Stockholms universitet

0. Inledning

Homografi eller lexikal flertydighet blir ett stort problem så snart man börjar försöka analysera större mängder text. Analys av verkliga texter från skilda genrer kräver ett generellt lexikon med hög täckningsgrad och för att den lexikala analysen ska kunna ligga till grund för annat än triviala lingvistiska iakttagelser krävs att antalet möjliga analysdistinktioner (t ex ordklassser, subkategorier eller morfologiska egenskaper) är stort. Båda dessa förutsättningar försvårar den lexikala analysen: Antalet homografer ökar!

Några triviala exempel:

När ett lexikons täckningsgrad ökar, ökar också chansen att en mindre frekvent tolkning av en ordform finns som alternativ i detta lexikon. Om det använda kategorisystemet skiljer mellan finita och infinita verbformer uppstår i många svenska verbböjningsmönster homografi mellan en preteritum- och en participtolkning.

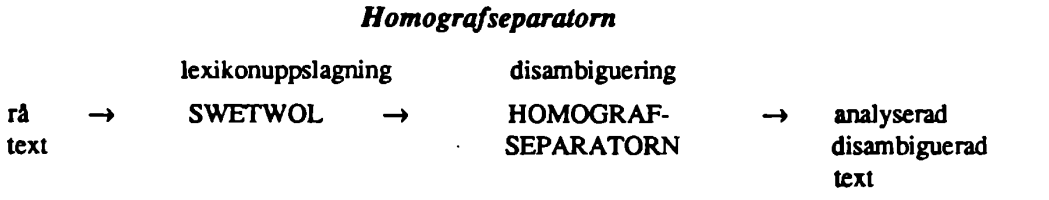
Redan vid ett ganska blygsamt antal analysdistinktioner kan graden av flertydighet bli besvärande. Som jag redovisar nedan ger ett lexikon med hög täckning och ett system av 11 kategorier (i huvudsak motsvarande de traditionella ordklasserna) till resultat att 36% av orden i de använda texterna är minst tvåfalt homografa. Om man utökar analyskategorierna till att omfatta även traditionella morfologiska drag kryper homografigraden för texterna upp över 50%.

I arbetet med att bygga upp Stockholm-Umeå Corpus, SUC, konfronterades vi snart med detta problem. Projektet (se Källgren 1990) har satt sig före att bygga upp en korpus motsvarande de engelska Brown- och LOB-korpusarna. Korpusen ska innehålla (minst) 1 milj. ord och dessa ska vara försedda med en (entydig) lexikal och morfologisk analys. I största möjliga utsträckning ska analysen utföras automatiskt - även om den första miljonen ord kommer att vara kontrollerad av mänskliga ögon - och för att nå fram till detta mål ska projektet även utvärdera olika metoder för automatisk lexikal disambiguering. Flera metoder och algoritmer ska testas, såväl lingvistiskt regelbaserade som sådana som baseras på frekvens hos tidigare analyserade korpusar. Här kommer att rapporteras om en pågående utvärdering av en metod av den senare typen.

1. Homografseparatorn

I fig. 1 nedan visas homografseparatorns plats i analysprocessen. En text tilldelas i det första steget alla möjliga analysalternativ av lexikon. I nästa steg filtreras de mindre sannolika alternativen bort av homografseparatorn.

Proceedings of NODALIDA 1991, pages 79-85



Figur 1

Det lexikon, SWETWOL, som projektet har fått möjlighet att använda baseras på Koskenniemis tvånivåmodell och är uppbyggt vid enheten för datorlingvistik vid institutionen för allmän språkvetenskap i Helsingfors (se Koskenniemi 1983 resp. Karlsson 1992).

1.1 Utgångsmaterial för statistik

För att kunna utarbeta en metod som ska baseras på tidigare förekomst i text behövs naturligtvis tillgång till tidigare analyserade texter. Jag har kunnat använda mig av Nusvensk frekvensordbok, NFO (Allén 1970), som innehåller frekvensuppgifter om 1 miljon löpord i Press 65-materialet och av Skrivsyntax, en mindre korpus bestående av texter från fyra olika genrer (se Teleman 1974). Under det fortsatta arbetet med SUC-korpusen kan de redan analyserade texterna i korpusen tillsammans med ovanstående material utgöra underlag för ny (och förbättrad) statistik.

Från dessa två källor har två sorters statistik hämtats. Från NFO har hämtats alla ord som enligt min kategoriuppsättning är homografa, 2404 ord. Från Skrivsyntax har jag hämtat statistik om samförekomst mellan kategorier. Skrivsyntax' korpus innehåller, i den form jag har använt den, 105 529 "lexikala enheter" dvs ord, lexikaliserade fraser och skiljetecken. I exemplen (1) och (2) nedan ges exempel på dessa olika typer av frekvenser.

(1)	*toppar*	verb pres	4
		substantiv pl	8
		totalt	12

(2)	substantiv – verb	4354
	pronomen – verb	3260
	substantiv – substantiv	1112

I (1) kan man alltså notera att ordformen *toppar* förekommer totalt 12 gånger i Press 65, därav 4 gånger som substantiv i plural, de övriga 8 gångerna som verb i presens. I (2) visas att ett verb föregås av ett substantiv i 4354 fall i Skrivsyntaxkorpusen, pronomen föregår verb i 3260 fall och ett substantiv föregås av ett annat substantiv i 1112 fall.

1.2 Övergångssannolikhet och lexikal sannolikhet

Processen kan översiktligt beskrivas på följande sätt: Homografseparatorn väljer den bästa tolkningen av en viss ordform genom att av den relativa frekvensen dra slutsatser om den mest sannolika tolkningen. Nedan är alltså nämnaren i exempel (3) den totala förekomsten av ordformen *toppar* och i (4) det totala antalet substantiv. De två typerna av information kombineras i (5) genom att de multipliceras samman och separatorn väljer den tolkning som har den högsta sammanslagna sannolikheten.

(3) lexikal sannolikhet (P-lex):

toppar	
V	4/12 = 0.33
N	8/12 = 0.67

(4) övergångssannolikhet (P-överg):

<entydigt ord>	<flertydigt ord>	
N	V	4354/21562 = 0.20
N	N	1112/21562 = 0.05

(5) kombination av lexikal sannolikhet och övergångssannolikhet:

<entydigt ord>	*toppar*	
N	V	0.33 * 0.20 = 0.066
N	N	0.67 * 0.05 = 0.034

1.3 Disambiguering av räckor av flertydiga ord

Val av den troligaste tolkningen för vart och ett av orden i en längre räcka, ett "spann" (se (6) nedan), av flertydiga ord kan i princip fortgå på liknande sätt. Sannolikhetsvärdet för varje tolkning beräknas som ovan och värdet för varje möjlig sekvens av analyser beräknas genom att de ingående alternativens värden multipliceras. Algoritmen beskrivs punktvis nedan.

(6) Ett spann innehållande N antal ord och kategoriema A-I. ... föreslagna av lexikon.

ord1	ord2	ord3	...	ord(N-1)	ordN
	B	D	...	G	
A	E		...		I
	C	F	...	H	

1) Bilda alla möjliga sekvenser av kategoriförslag för spannet:

A	B	D	...	G	I
A	B	D	...	H	I
A	B	E	...	G	I
...			...		
A	C	F	...	H	I

2) Beräkna sannolikhetsvärdet för varje sekvens genom att multiplicera de enskilda alternativens. Nedan visas alternativsekvensen A-B-D-...-G-I.

ord1	A	
ord2	B	[P-lex(B,ord2) * P-överg(A,B)] *
ord3	D	[P-lex(D,ord3) * P-överg(B,D)] *
...	...	[...] *
ord(N-1)	G	[P-lex(G,ordN-1) * P-överg(G,...)] *
ordNI		[P-överg(G,I)] = sannolikheten för kombination 1.

3) Välj kombinationen med det högsta sannolikhetsvärdet.

En algoritm som denna är mycket resurskrävande. Homografseparatorn använder en annan algoritm presenterad i DeRose 1988. Hans algoritm, VOLSUNGA, bygger på "dynamisk programmering" och löser problemet mer effektivt. Den baseras på följande iakttagelse: Om ett visst analysalternativ ingår i den bästa sekvensen av alternativ för hela spannet måste även den bästa delsekvens som föregår analysalternativet ingå i den bästa sekvensen. Detta innebär att man för varje tolkningsalternativ bara behöver spara den bästa sekvensen av tolkningar fram till detta

alternativ. I spannet i (6) ovan innebär användningen av DeRoses algoritm att vid processningen av ord4 endast två olika kombinationer av tolkningar (nämligen den bästa vägen till D och den bästa vägen till F) behöver analyseras för vart och ett av detta ords alternativ. Detta ska jämföras med de sex sekvenser som den primitivare algoritmen måste hantera.

2. *Utvärdering*

En utvärdering av denna metod för lexikal disambiguering har påbörjats. Utvärderingen ska baseras på homografseparatorns analys av en större mängd texter. Dessa texter är helt skilda från de texter som utgör bas för den statistik som separatorn använder. I denna artikel utvärderas analysen av ett mycket litet antal ord, 3103 st, fördelade på två olika texter.

Den kategoriuppsättning som har använts är begränsad och innehåller 11 kategorier som ungefärligt motsvarar de traditionella ordklasserna och 3 kategorier för olika typer av skiljetecken. Detta innebär som tidigare nämnts att homografseparatorn ges en enklare uppgift jämfört med disambigueringen av alternativ från ett mer finindelande system. Det finns två skäl till detta. Det första är rent praktiskt: De använda kategorierna är skärningsmängden av de kategorier som används av SWETWOL, NFO och Skrivsyntax. Det andra skälet är att vi ville undersöka hur mycket ett så pass grovmaskigt nät kunde sålla bland alternativen i ett mer omfattande system. Kategorierna är alltså inte desamma som används för analys av SUC-korpusen. Fig 2 visar kategorisystemet och fig. 3 resultaten av denna första utvärdering.

Figur 2 Använda kategorier

substantiv	N	konjunktion	KONJ
verb	V	preposition	PREP
adjektiv	A	pronomen	PRON
adverb	ADV	räkneord	NUM
interjektion	INTERJ	infinitivmärke	INFMARK
egennamn	<Prop>		

och

"stora" skiljetecken (.?!:)	CB
komma (,)	IK
övriga skiljetecken	IG

Figur 3 Resultat (3103 ord)

instanser av ambiguitet / totalt antal ord	1109/3103	36%
korrekt analys / ambiguiteter	974/1109	87%
felaktig analys / totalt antal ord	135/3103	4%
kvarvarande ambiguitet		0%

Man kan notera att homografseparatorn vid analys av de ca 3100 orden väljer en riktig analys i 87% av de fall den utsätts för. Till detta kommer 1994 ord som får en entydig analys av lexikon och där homografseparatorn alltså inte appliceras. Av dessa får ett fåtal en felaktig lexikonanalys men totalt sett är ändå närmare 96% av orden i texten korrekt analyserade efter separatorns arbete.

2.1 Några exempel på analyser

Nedan redovisar jag några exempel från analysen. A-exemplen visar det flertydiga spannet före disambigueringen, b- och c-exemplen vilken analys homografseparatorn valt.

I (7a) visas ett spann som innehåller 7 flertydiga ordformer i följd och de sammanlagt 288 möjliga analyserna för spannet. (7b) visar homografseparatorns korrekta analys.

(7a)

iakttar	*vad*	*som*	*händer*	*på*	*andra*	*håll*	*i*	*världen*
V	ADV	KONJ	V	PREP	NUM	N	PREP	N
	N	PRON	N	ADV	PRON	V	ADV	
	PRON				V			

(7b)

iakttar	*vad*	*som*	*händer*	*på*	*andra*	*håll*	*i*	*världen*
V	PRON	PRON	V	PREP	PRON	N	PREP	N

I (8a) kan man notera att en ordform, *bedragnade*, inte ges något förslag till analys av SWETWOL-lexikonet. Homografseparatorn hanterar av lexikonet oanalyserade ord på enklast tänkbara sätt. En ordform som saknar analysförslag görs av homografseparatorn till ambiguöst mellan alla kategorier i det använda kategorisystemet, f.n. 14 stycken. Detta innebär alltså att disambigueringsprocessen i detta fall ska ta ställning till 112 möjliga analyser för spannet. (8b) visar den (nästan) korrekta analysen. Strikt bedömt skulle *bedragnade* ha analyserats som V enligt mitt kategorisystem men jag väljer att godkänna även adjektivanalysen eftersom att bestämma perfekt participformers ordklasstillhörighet inte är ett problem bara för homografseparatorn!

(8a)

ej	*bedragnade*	*band*	*av*	*dubbla*	*löften*
ADV		V	PREP	V	N
		N	ADV	A	

(8b)

ej	*bedragnade*	*band*	*av*	*dubblal*	*öften*
ADV	A	N	PREP	A	N

(9) är exempel på en typisk svaghet hos separatorn. Att avgöra om *som* ska analyseras som KONJ eller PRON klarar den sällan!

(9a)

världen	*som*	*blott*	*en*	*illusion*
N	KONJ	ADV	ADV	N
	PRON	KONJ	N	
		A	PRON	

(9b)

världen	*som*	*blott*	*en*	*illusion*
N	KONJ	ADV	PRON	N

Som nämnts hanteras oanalyserade ordformer mycket enkelt. För att förbättra homografseparatorns prestation är det möjligt att använda sig av ett flertal heuristiska regler av typen: identifiering av ändelser, användning i texten av versaler/gemena, etc. Ingenting av detta är implementerat i den nuvarande versionen av homografseparatorn. C-exemplen nedan visar att effektiviteten kan förbättras genom att bara begränsa de kategorier som kan tillskrivas ett oanalyserat ord. I dessa exempel har separatorn bara bedömt sannolikheten för att ett oanalyserat ord ska tillhöra någon av de öppna kategorierna: N, A, V, <Prop> eller ADV, medan b-exemplen väljer bland 14 kategorier inklusive skiljeteckenkategorier som CB (clause boundary).

(11a)

.	*Nibelungen*	*Alberich*	*rövar*	*rhenguldet*	*och*
CB			V		KONJ
			N		

(11b)

.	*Nibelungen*	*Alberich*	*rövar*	*rhenguldet*	*och*
CB	N	PREP	N	CB	KONJ

(11c)

.	*Nibelungen*	*Alberich*	*rövar*	*rhenguldeto*	*ch*
CB	N	V	V	N	KONJ

(12a)

.	*Nibelungarnas*	*fångenskap*
CB		N

(12b)

.	*Nibelungarnas*	*fångenskap*
CB	PRON	N

(12c)

.	*Nibelungarnas*	*fångenskap*
CB	N	N

3. Sammanfattning

Redan dessa preliminära försök visar att man kan nå överraskande goda resultat med denna enkla metod. Vi är medvetna om att när vi gradvis utökar antalet kategorier kommer de goda resultaten att försämras men samtidigt kommer korpusen ständigt att växa vilket ger oss ett växande material att basera och därmed förbättra statistiken på. Det allt större materialet ger oss också möjlighet att pröva mer sofistikerade metoder för statistiskt baserad disambiguering. Vi bedömer det vara fullt möjligt att uppnå 96-97% ordklasskorrekthet även med ett utökat kategorisystem, men hur mycket längre kan man komma?

Det har, oss veterligt, heller ännu inete någonstans, gjorts några försök att disambiguera morfologiska särdrag med samma metoder som här har använts för disambiguering av ordklasskategorier. Sådana försök står vi i begrepp att utföra inom SUC-projektets ramar.

Referenser

Allén, S., 1970. *Nusvensk frekvensordbok.* Almqvist & Wiksell. Stockholm.

DeRose, S.,1988. *Grammatical Category Disambiguation by Statistical Optimization.* Computational Linguistics, 14,1.

Karlsson, F., 1992. *SWETWOL: A Comprehensive Morphological Analyzer for Swedish.* Nordic Journal of Linguistics 1:1992.

Koskenniemi, K., 1983. *Two-level Morphology: A General Computational Model for Word-form Recognition and Production.* Publications of the Department of General Linguistics, University of Helsinki, No. 11.

Källgren, G. 1990. 'The first million is hardest to get'. COLING 1991.

Teleman, U., 1974. *Manual för grammatisk beskrivning av talad och skriven svenska.* Studentlitteratur. Lund.

Gunnar Eriksson
Institutionen för lingvistik
Stockholms universitet
S-106 91 Stockholm
E-mail: gunnar@ling.su.se