

# Flerspråkig informationsextraktion

Stina Nylander

maj 2000

## Inledning

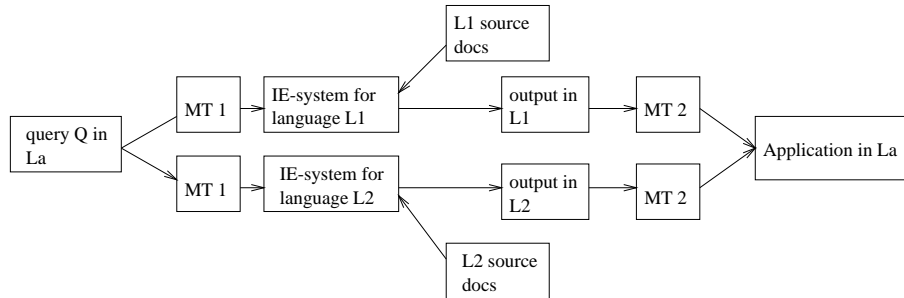
Informationsextraktion är en disciplin som till skillnad från Information Retrieval (IR) (Dini, 1998) varit enspråkig under större delen av sin livstid. Detta har varit fallet även de gånger när MUC-konferenserna ställt upp flerspråkiga problem; de deltagande grupperna har utvecklat separata system för varje inblandat språk (Gaizauskas et al., 1997). Målet med flerspråkig informationsextraktion är att samma IE-system ska kunna extrahera information ur material på flera språk och presentera resultatet på ett språk som användaren förstår. Dagens system är så pass tätt integrerade med den fråga de ska söka svar till att en funktion som ger möjlighet att fråga på olika språk inte är aktuell. I framtiden vill man dock att systemen ska bli så lätta att anpassa till nya problem och områden att även det språk frågan ställs på bör bli anpassningsbart.

För att kunna genomföra flerspråkig informationsextraktion behöver man överbrygga skillnaderna mellan de inblandade språken, och de metoder som förs fram i det material jag har läst handlar om olika typer av språkoberoende representationer av textinnehåll, eller om maskinöversättning. Jag har läst Kameyamas artikel om en tänkbar design av ett IE-system och artiklar om två prototyper, MIETTA och M-LaSIE, vilka alla presenteras mer i detalj nedan.

## IE och MT

Kameyama har inte utvecklat något extraktionssystem, men presenterar i sin artikel *Concepticons vs. Lexicons: An Architecture for Multilingual Information Extraction* (1997) en tänkbar arkitektur (se fig 1). Enligt Kameyama är första steget mot flerspråkig informationsextraktion att göra dagens system mer flexibla (*open target*), vilket gör att det språk man ställer frågan på också tas med i resonemanget. Kameyama föreslår en systemdesign med parallella IE-system, ett för varje enskilt språk, med maskinöversättningssystem som brygga mellan språken. Frågan översätts till de olika ingående språken, IE-system för varje språk extraherar relevant information och svaren översätts till det språk som användaren föredrar.

Fördelarna med detta kan vara att det är lättare att få fram existerande maskinöversättningssystem än system som skapar språkoberoende textrepre-



Figur 1: Kameyamas designförslag

sentationer och att ifyllda templat är ett textmaterial som lämpar sig väl för maskinöversättning. Nackdelar kan vara att både informationsextraktion och maskinöversättning är domänberoende vilket kan försvåra domänbyte, och att viss information kan förvrängas när både frågan och svaret översätts.

## MIETTA

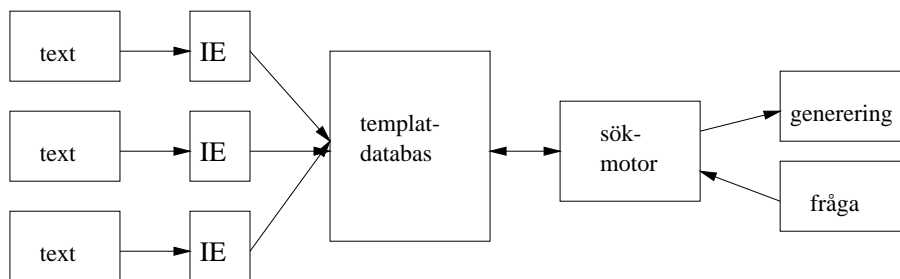
MIETTA står för *Multilingual Information Extraction for Tourism and Travel Assistance* och är precis som den heter en prototyp för extraktion av turistinformation som ska fungera för fyra språk, engelska, finska, italienska och tyska. Tanken är att den framtida användaren ska kunna använda till exempel Netscape som gränssnitt till MIETTA:s server. Projektet har en hemsida på <http://mietta.dfki.de>.

MIETTA bygger på parallella IE-system som extraherar information till språkoberoende templat som lagras i en databas. Till databasen kan sedan frågor ställas i naturligt språk, och från de templat som matchar frågan genereras svaret på användarens språk (se fig 2). Eftersom de ifylla templaterna lagras i en databas behöver extraktionen bara genomföras en gång för varje text, och belastar bara systemet när nya texter läggs till. Man utgår ifrån att det textmaterial som är intressant kommer att vara relativt oföränderligt och att samma ifyllda templat kommer att behövas flera gånger. Dini (1998) trycker hårt på att templaterna i ett IE-system måste fyllas i på ett språkoberoende sätt, det vill säga med definierade attribut, så att en ifylld templat aldrig innehåller en för systemet okänd sträng. En uttalad parallell dras till maskinöversättningsområdet interlingua. Utifrån de ifyllda templaterna kan sedan svaret genereras på det språk användaren föredras, utan att någon maskinöversättning behövs.

Ingen av de två artiklar (Buitelaar, 1998; Dini, 1998) jag har hittat om MIETTA redovisar några resultat.

En fördel med denna systemdesign är att frågan inte riskerar att förvrängas vid översättning och inte heller svaret eftersom det genereras automatiskt. En nackdel är att den som bygger systemet avgör vilka templat som skall extraheras till databasen vilket gör att den kan fyllas av templat som ingen

## MIETTA



Figur 2: MIETTA:s uppbyggnad

någonsin frågar efter. Enligt Dini (1998) skulle de språkoberoende templaterna drastiskt kunna minska det arbete som nu måste läggas ned på underhåll av flerspråkiga databaser.

## M-LaSIE

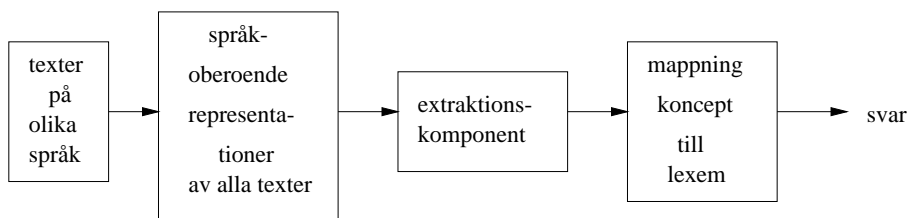
M-LaSIE är en prototyp för informationsextraktion från fransk och engelsk text som bygger på det enbart engelskspråkiga extraktionssystemet LaSIE, utvecklat vid universitetet i Sheffield.

Både M-LaSIE och MIETTA använder sig av språkoberoende representation av texter, men skiljer sig tydligt på andra punkter. M-LaSIE börjar med att bygga upp en språkoberoende representation av texten, en så kallad diskursmodell (Gaizauskas et al., 1997), ur vilket sedan önskad information extraheras. Diskursmodellen byggs upp genom att varje mening först översätts till en kvasi-logisk form; till detta läggs sedan ytterligare presuppositioner och koreferenslösning. Denna diskursmodell blir då genomsam för alla språk, vilket gör att man kan använda samma extraktionsmodell för alla ingående språk. Svaren skapas genom att de extraherade koncepten matchas mot svarsspråkets lexikon (se fig 3). Gaizauskas et al. (1997) trycker hårt på att man lyckats separera konceptuell och lexikalisk kunskap helt i systemet, och på detta sätt gjort det enkelt att lägga till ytterligare ett språk. Man behöver då endast göra ett lexikon för språket i fråga och koppla detta till koncepten i diskursmodellen.

Gaizauskas et al. (1997) tar trots sina språkoberoende representationer avstånd från det generella *interlingua* man ibland talar om inom maskinöversättningen, men säger att man tror att det är möjligt att skapa språkoberoende representationer för specifika domäner.

M-LaSIE har använt textmaterial från MUC 6 och testats på 20 korta parallella texter. Resultatet blev likadant för både engelska och franska, dvs samma händelser plockades fram ur texterna.

Fördelen med den systemdesign man valt för M-LaSIE är att man kan



Figur 3: M-LaSIE:s uppbyggnad

använda samma extraktionsmodul för alla språk, man behöver dock skapa relationer mellan lexikon och diskursmodell för varje språk. Nackdelen är att man skapar en språkoberoende representation av en hel text för att sedan extrahera det man är intresserad av. Frågan är om det inte kräver så mycket jobb för att skapa den språkoberoende representationen att man inte vinner så mycket på att använda ett enda extraktionssystem.

## Diskussion

Flerspråkig informationsextraktion är ett ganska nytt forskningsområde, som ännu inte riktigt kommit igång. De system som finns är i högsta grad under utveckling och mycket prototypiska.

Den litteratur jag har läst (som på intet sätt gör anspråk på att vara heltäckande för området) visar att det ännu inte verkar ha skett samma likriktning vad gäller systemens uppbyggnad som man kan se i de enspråkiga IE-systemen. Även om Kamayama inte bygger ett system utifrån sina tankar om att kombinera IE och maskinöversättning är det ett designförslag som skiljer sig markant från de andra två systemen jag tittat på. MIETTA och M-LaSIE liknar varandra så tillvida att de båda använder sig av en språkoberoende representation, med de använder den på olika sätt. Det råder även delade meningar om hur många IE-system man ska använda, ett eller flera?

Ytterligare ett sätt att skapa flerspråkig IE skulle kunna vara det Kristoffer föreslog på seminariet, nämligen att översätta extraktionsreglerna från ett enspråkigt IE-system.

## Litteratur

Buitelaar, P., Netter, K., Xu, F. 1998. Integrating Different Strategies in Cross-Language Information Retrieval in the MIETTA Project, Proceedings of TWLT14, Enschede, the Netherlands.

Dini, L. 1998. Parallel IE systems for multilingual information gathering. Euriscon 98, Third European Robotics, Intelligent Systems & Control Conference.

Gaizauskas, R., Humphreys, K., Azzam, S., Wilks, Y. 1997. Concepticons vs. Lexicons: An Architecture for Multilingual Information Extraction. In Maria Teresa Pazienza, editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, volume 1299 of *Lecture Notes in Artificial Intelligence*, chapter 3, pp. 28-43. Springer. International Summer School, SCIE-97. Frascati, Italy.

Kameyama, M. 1997. Information Extraction across Linguistic Barriers. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, Stanford University, March 24-26.