Namngivna entiteter – en svensk guldstandard och utvärderingsresurs

Lars Ahrenberg, Johan Frid och Leif-Jöran Olsson

lars.ahrenberg@liu.se, johan.frid@humlab.lu.se,
leif-joran.olsson@svenska.gu.se

Bakgrund

Den mest använda svenska guldstandarden för svenska är SUC-3, som dock inte speglar aktuellt svenskt språkbruk. Den omfattar många namn, men med ojämn fördelning över entitetsklasser.
De flesta infrastrukturprojekt som skapar språkresurser idag behöver märka upp namngivna entiter, t.ex. för syften som sökning, metadataupmärkning eller pseudonymisering.
Swe-Clarin har därför som ett av sina temaområden för 2018-19 valt namngivna entiteter (named entities).
Ett exempel på en specifik ingång kommer från medicinskt håll, där man intresserar sig för hur informationsextrahering ur olika typer av textmaterial kan förbättra patienthandläggning. Det kan t ex röra sig om att identifiera uppgifter i patientjournaler som tidigare sjukdomar, vilka slags tester som tagits och pågående behandlingar. Identifierade enheter kan sedan användas i olika applikationer, t ex beslutstödssystem.

Syfte

Syftet är att ta fram en svensk guldstandard som kan användas för benchmarking och som lätt kan utvidgas. Definitioner och principer för uppmärkning ska dokumenteras.

Metod/Genomförande

Iterativt arbete

- åtta valda kategorier:
  - personer, yrken
  - platser (geografiska platser)
  - tidsangivelser
  - organisationer
  - works/artefakter/produkter
  - händelser
  - symptom (allergier)
  - behandling (mediciner, drugs)
- 1000 instanser
- enkel atomär uppmärkning (en nivå) till att börja med t ex *PRS* för kategorin personer. När vi i framtiden ev inför undertypning kan detta markeras med t ex *LOC-CTY* för namnet *Stockholm*.
- successiv utveckling av riktlinjer

I annoteringssteget är materialet grunduppmärkt ut på detta sätt med token, ev existerande NER-tagg, och ordklass:

	A	B	C
1	Viggo	PRS	PM
2	var	O	VB
3	sugen	O	PC
4	på	O	PP
5	makaroner	O	NN
6	å	O	KN
7	korv	O	NN
8	å	O	KN
9	det	O	PN
10	ser	O	VB
11	jag	O	PN
12	verkligen	O	AB
13	som	O	KN
14	ett	O	DT
15	gott	O	JJ
16	tecken	O	NN
17	.	O	MAD
18	:)	O	LE
19			
20	Det	O	PN

Preliminära generella riktlinjer för taggning

- Riktlinjerna uppdateras med jämna mellanrum utifrån vunna erfarenheter.
- Riktlinjerna tillämpas på de texter som vi vill använda för träning eller testning av olika NERC-system för svenska.
- Ett annat mål är att ge vägledning för hur de framtida riktlinjerna för anonymisering ska utformas.

Taggningsformat

- En enskild tagg har formatet *P-TYP*, där *P* anger position i ett namnuttryck och *TYP* anger vilken typ av entitet som namnet refererar till.
- *P* kan vara antingen *B* eller *I*.
  - *B* används för det första ordet i ett namnuttryck
  - *I* för alla följande ord, om namnuttrycket består av flera ord.

TYP kan vara något av följande för namnuttryck:

- *PRS* refererar till en person
- *LOC* refererar till en plats eller annat geografiskt område
- *ORG* refererar till en arbetsplats, skola eller liknande
- *TMP* om namnuttrycket är en tidsangivelse
- *EVT* anger en specifik händelse, eller serie händelser som på grund av sin betydelse fått ett specifikt namn
- *TCUDORP* refererar till en artefakt, produkt eller konstnärligt verk
- *SYMP* refererar till ett symptom av sjukdom
- *MEDTREAT* refererar till en behandling eller medicinering

Avgränsning av typer

- *Personer (PRS)*. Alla personnamn märks upp. Även personnummer märks upp.
  - Om en titel finns angiven i direkt anslutning till ett personnamn antas detta ingå i namnet. Exempel: *apotekare Lundin*.
  - Även initialer märks upp.
  - Nära relation-utryck bör också markeras. Exempel: *min man*, *chefen*.
  - Personliga pronomen som *hon* och *han* markeras inte.
- *Platser (LOC)*. Namn på länder, län, landskap, regioner, städer, byar och stadsdelar. Även namn på gator, parker och annat som kan identifiera en bostads- eller arbetsort märks upp.
  - Beskrivande substantiv kan ingå i namn av typen *Östergötlands län*, *Norrlands inland*.
- *Organisationer (ORG)*. Namn på företag, skolor, sjukhus, myndigheter, föreningar, politiska partier markeras. Även utpekande namn på avdelningar eller sektioner markeras. Här finns dock gränsdragningsproblem och **den generella rekommendationen är att hellre annotera för mycket än för lite**.
- *Tidsangivelser (TMP)*. Här markeras uttryck som anger specifika år, månader och datum. Även tidsintervall som *1970-75* eller *60-talet* markeras. Vi markerar tidsuttryck som hela fraser.
- *Händelser (EVT)*. Här markeras uttryck som markerar viktiga händelser där en person kunnat delta. Exempel *Andra världskriget*, *Socialdemokraternas partistämma*, *Apoteksbolagets miljökonferens*.

Data/Material

Urvalet är gjort för att fånga modernt språk i texter från olika genrer. Vårt material utgörs av två huvuddelar. En del med icke-meningsomkastade utsnitt och en del med meningsomkastade utsnitt. Meningsomkastningen är nödvändig på grund av copyright-skäl. Annars skulle spridningen över genrer inte vara så stor.

- Meningsomkastat
  - Flashback-fordon 2010
  - Familjeliv-barnhälsa 2010
  - GP 2010
  - Bloggmix 2010
- Ej meningsomkastat
  - Wikipedia
  - Stockholm Internet Corpus (SIC)
- Partitionerat för att få mått på interbedömar-reliabilitet
  - 20 delar per utsnitt
  - LibreOffice Online
- Resursen öppen för att arbeta vidare med

Benchmarking

I första hand vill vi använda befintliga system, t ex i Sparv, Stanford-corenlp, etc.

```
5  let $lang := "swe"
6  let $lang-map := map {"eng": 1, "swe": 2, "ces": 3}
7  let $classifier := (xs:anyURI("/db/apps/stanford-ner/resources/classifiers/english.all.3class.distsim
8  (:    :xs:anyURI("/db/apps/stanford-ner/resources/classifiers/swe-outfile-300dpi-2class-model.ser.gz")
9  xs:anyURI("/db/apps/stanford-ner/resources/classifiers/swe-dw12-3class-model.ser.gz"),
10 xs:anyURI("/db/apps/stanford-ner/resources/classifiers/ces-1872-outfile-djvu-2class-model.ser.gz"))[x
   ?($lang)]
11 let $text := (<p>The fate of Lehman Brothers, the beleaguered investment bank,
12    hung in the balance on Sunday as Federal Reserve officials and the leaders
13    of major financial institutions continued to gather in emergency meetings
14    trying to complete a plan to rescue the stricken bank.  Several possible
15    plans emerged from the talks, held at the Federal Reserve Bank of New York
```

/db/temp/stanford-test.xquery

XML Output ▾ Live Preview

```
<p>När det gäller
   <pers>Åke</pers>
och hans värld kommer vi långt ifrån bank- och finansväsendet.
   <pers>Mimmi</pers>
och Sonja är i fjällen på semester när lavinen går på
   <org>Blåsjöfjället</org>
. Skårgårdens verklighet en vinterdag är inte heller så rosenskimrande. Herr
   <pers>Arne</pers>
är dock i nya världen för gull och penningar. Ute på ön är oron stor för
   <pers>Olagus</pers>
```

Vi vill även i uppdelningen meningsomkastat/icke-meningsomkastat se om något resultatmått påverkas av om materialet är löpande med större kontext eller endast lokal kontext på meningsnivå i det meningsomkastade.