# Övervakad namntaggning med domänspecifik träningsdata

Adam Persson

Institutionen för lingvistik
Datorlingvistik
Examensarbete för kandidatexamen 15 hp
Kandidatprogram i lingvistik 180 hp
Vårterminen 2016
Handledare: Kristina Nilsson Björkenstam
Examinator: Bernhard Wälchli
Expertgranskare: Mats Wirén
English title: Supervised named-entity recognition with domain-specific training data

Stockholms universitet

# Övervakad namntaggning med domänspecifik träningsdata

**Adam Persson**

# Sammanfattning

Övervakad maskininlärning har gett goda resultat för automatisk namntaggning. Detta kräver dock manuellt annoterad träningsdata, vilket är krävande att ta fram. Studier har visat att likhet mellan träningsdata och testdata är viktigt för att uppnå bra resultat, men normalt sett tränas system alltid med så mycket data som möjligt, utan hänsyn till dess relevans. Syftet med denna studie är att undersöka om bättre namntaggning kan uppnås genom att utesluta de delar av träningsdatan som inte tillhör samma textdomän som testdatan. För att genomföra detta konstrueras ett system med multinomial logistisk regression som tränas och testas på Stockholm-Umeå Corpus enligt både traditionell och föreslagen metod. Undersökningen visar en liten men signifikant försämring vid användning av enbart domänspecifik träningsdata, ett resultat som dock inte är genomgående för alla delar av undersökningen. Den stora fördelen av att reducera träningsdatan är dock att det ökar maskininlärningens hastighet. För att kunna utnyttja detta föreslås att namntaggning föregås av textklassificering.

### Nyckelord

Namntaggning, övervakad maskininlärning, multinomial logistisk regression, domänspecifik träningsdata

# Supervised named-entity recognition with domain-specific training data

**Adam Persson**

# Abstract

Supervised machine learning has given good results for automatic named-entity recognition. However, this requires manually annotated training data, which is demanding to produce. Studies have shown that similarity between training data and test data is important for good results, but systems are normally trained with all available data, ignoring its relevance. The purpose of this study is to investigate if named-entity recognition can be improved by reducing training data by removing text that does not belong to the same domain as the test data. To do this, a system using multinomial logistic regression is constructed and evalutated on Stockholm-Umeå Corpus with both the traditional and the suggested method. The main results show a small but significant deterioration when only using domain-specific training data, but this result doesn't hold true for every part of the experiment. The major benefit of reducing training data is however the improvement in computing time. To truly be able to take advantage of this, it is proposed that named-entity recognition is preceded by text classification.

### Keywords

Named-entity recognition, supervised machine learning, multinomial logistic regression, domain-specific training data

# Innehållsförteckning

**1 Inledning ..... 1**
**2 Bakgrund ..... 2**
2.1 Begrepp ..... 2
2.1.1 Namn ..... 2
2.1.2 Domän ..... 2
2.2 Kort ämnesbakgrund ..... 2
2.3 Konstruktion av namntaggningssytem ..... 3
2.3.1 Övervakad maskininlärning ..... 3
2.3.2 Namnlistor ..... 4
2.3.3 Taggset ..... 4
2.4 Utvärdering av namntaggningssystem ..... 5
2.4.1 Precision, täckning, och F1-värde ..... 5
2.4.2 K-faldig korsvalidering ..... 5
2.4.3 Problematik med utvärdering ..... 6
2.5 Stockholm-Umeå Corpus ..... 6
2.6 Namntaggningssystem för svenska ..... 7
**3 Syfte och frågeställningar ..... 8**
3.1 Syfte ..... 8
3.2 Frågeställningar ..... 8
**4 Metod ..... 9**
4.1 Systemarkitektur ..... 9
4.1.1 Träningsdata och särdrag ..... 9
4.1.2 Framtagna namnlistor ..... 10
4.1.3 Sekventiell namntaggning ..... 10
4.2 Utvärdering av systemet ..... 10
4.2.1 Jämförelse av namnmängder ..... 10
4.2.2 Korsvalidering ..... 11
4.3 Experimentdesign ..... 11
**5 Resultat ..... 12**
5.1 Experimentresultat ..... 12
5.1.1 Resultat utifrån frågeställningar ..... 12
5.1.2 Fullständigt korrekt taggade namn ..... 12
5.1.3 Åtminstone delvis korrekt taggade namn ..... 13
5.2 Systemets utvärderingsresultat ..... 14
**6 Diskussion ..... 15**
6.1 Metoddiskussion ..... 15
6.1.1 Diskussion om maskininlärningsmetod ..... 15

6.1.2 Diskussion om Stockholm-Umeå Corpus ..... 15
6.1.3 Diskussion om använt taggset ..... 15
6.1.4 Diskussion om använda namnlistor ..... 16
6.1.5 Diskussion om experimentets utformning ..... 16
6.2 Resultatdiskussion ..... 17
6.2.1 Diskussion om huvudresultat ..... 17
6.2.2 Diskussion om resultat utifrån namnklasser ..... 17
6.2.3 Diskussion om resultat utifrån textdomäner ..... 17
**7 Slutsatser ..... 18**
**Referenser ..... 19**
**Bilagor ..... 21**
A: SUC-domäner ..... 21
B: Signifikanstester ..... 22

# 1 Inledning

Namntaggning är en klassificeringsuppgift inom datorlingvistik och informationsutvinning där målet är att identifiera namn i texter och tagga (klassificera) dem enligt ett på förhand bestämt taggset (en mängd klasser). Med språkteknologiska metoder kan detta göras automatiskt, till en viss grad av korrekthet. Ett exempel på en namntaggad mening är ”Filip[person] bor i Danmark[plats]”.

Förutom att namntaggning kan bidra till bland annat förbättrad maskinöversättning och taligenkänning har det även användningsområden inom andra fält. Till exempel har Elson et al. (2010) använt namntaggning för att generera modeller av de sociala nätverk som uppstår mellan karaktärer i brittisk litteratur för att undersöka skillnader i hur det sociala livet i stadsmiljö och lantmiljö framställdes på 1800-talet i Storbritannien, baserat på vilka karaktärer som talar med varandra. Detta möjliggjordes på stor skala med hjälp av bland annat namntaggning.

En av de metoder som har använts för automatisk namntaggning är att träna en maskininlärningsmodell med manuellt namntaggad text, vilket är en typ av övervakad maskininlärning. Den text som används för att träna upp ett system bör då vara representativ för de texter som systemet ska namntagga, åtminstone i de aspekter som maskininlärningen baserar sina beslut på. Detta kan eventuellt uppnås genom att endast träna och använda en modell med texter från en och samma textdomän, men traditionellt sett tränar man sin modell med så mycket data som möjligt.

I denna studie konstrueras ett övervakat namntaggningssytem som använder multinomial logistisk regression. Systemet tränas och testas med Stockholm-Umeå Corpus. I den experimentella delen av studien utför detta system en namntaggningsuppgift två gånger, där skillnaden mellan de två körningarna är vilken träningsdata som används. I den första körningen används all tillgänglig träningsdata. I den andra körningen utesluts den träningsdata som inte tillhör samma textdomän som testdatan.

# 2 Bakgrund

## 2.1 Begrepp

### 2.1.1 Namn

Med namn avses inom namntaggning de explicita namn som refererar till en entitet. Detta exkluderar pronomen och beskrivningar som också kan referera till en entitet. Explicita namn kan dock se ut på olika sätt. Namn på personer kan till exempel bestå av endast ett token när de bara består av ett förnamn eller efternamn. Ett personnamn kan också bestå av titel, förnamn, mellannamn, initial och efternamn. Oavsett hur många token ett namn består av räknas det bara som ett namn. Verk som böcker och filmer kan ha namn som består av hela meningar.

### 2.1.2 Domän

I denna uppsats används ibland termen domän, och ibland textdomän. De syftar båda på en mängd av manuellt kategoriserade texter med någon meningsfull gemensam egenskap. Det råder förvirring över vad denna gemensamma egenskap är inom en domän, och hur domäner skiljer sig från genrer. Ett sätt att komma förbi detta problem är att istället dela in texter i så kallade textkategorier, en term som kan motsvara vilken sorts indelning som helst. Denna studie kommer att behandla textkategorier från en korpus som har använt denna neutrala lösning, men de kommer benämnas som domäner (eller textdomäner). För en längre diskussion om textkategorisering och terminologiproblematik inom området, se Lee, D. Y. (2001).

## 2.2 Kort ämnesbakgrund

Maskinell namntaggning var ett nytt koncept under första halvan av 1990-talet. En namntaggningsuppgift inkluderades i Message Understanding Conference-6 (MUC6), vilket kan ses som namntaggningens startskott. Uppgiften bestod av att tagga person- och organisationsnamn i texter från tidningen Wall Street Journal. Mycket bra resultat uppmättes. Majoriteten av de deltagande systemen utförde uppgiften med ett F1-värde (se avsnitt 2.4.1) över 90%. En viktig faktor till dessa goda resultat anses vara den ickevarierade och förutsägbara textmängd som Wall Street Journal-artiklarna utgjorde (Grishman, 1996).

Vid konferensen CoNLL-2003 bestod en uppgift av att skapa ett språkoberoende namntaggningssystem som skulle användas på både engelska och tyska. Systemen skulle identifiera fyra klasser av namn: personer, platser, organisationer, och övriga namn som inte passade in i någon av de andra klasserna. Den engelska datamängden används ofta fortfarande som jämförbart utvärderingsmaterial för engelskspråkiga namntaggare. Träningsdatan består av nyhetstexter från augusti 1996 och testdatan består av nyhetstexter från december 1996 (Tjong & De Meulder, 2003).

I forskningsprogrammet Automatic Content Extraction (ACE) ville man ta uppgiften ett steg längre genom att inte bara tagga de entiteter som finns i text som explicita namn, utan även pronomen och

beskrivningar. Eftersom detta inte bara omfattar namn är det inte längre namntaggning, utan entitetstaggning, vilket är en annan – och svårare – uppgift (Doddington et al., 2004).

Ratinov och Roth (2009) undersökte effekten av ett antal faktorer vid namntaggning och utvecklade ett system som gav det bästa publicerade resultatet på den engelska delen av testdatan från CoNLL-2003. Deras system testades även på en uppsättning webbsidor som de själva tagit fram, där texterna var av mycket mer varierande natur än den engelska delen av CoNLL-2003, som endast består av nyhetstexter. På grund av svårigheter att komma överens om hur vissa namn skulle taggas gjordes testet endast för namnidentifikation på tokennivå, utan att ta hänsyn till olika typer av namnklasser. Trots detta i jämförelse enkla test presterade systemet avsevärt mycket sämre när det tränades och testades med webbtexter. Detta illustrerar vikten av att träningsdatan liknar testdatan.

# 2.3 Konstruktion av namntaggningssytem

## 2.3.1 Övervakad maskininlärning

Maskininlärning handlar om att skapa system som kan förbättra sin förmåga att generalisera och fatta beslut baserat på erfarenheter, vilket möjliggörs av datorers processorkraft och diverse statistiska metoder för att hitta mönster i data. Övervakad maskininlärning innefattar de system som tränas med uppmärkt data, till exempel en text där varje token är annoterat med ordklass, eller information om huruvida det är en del av ett namn eller inte. Dessa uppmärkningar kallas responsvärden. En modell som tränats med uppmärkt data ska sedan kunna märka upp ny data genom att i vissa fördefinierade aspekter jämföra den med träningsdatan och bestämma vilket responsvärde som är mest sannolikt givet träningsdatan. Detta sker under antagandet att träningsdatan och testdatan delar de egenskaper som systemet kan ta till vara på. Vilka dessa egenskaper är beror på uppgiften (Kotsiantis et al. 2007).

Multinomial logistisk regression är en typ av övervakad maskininlärning som används för klassificeringsuppgifter där en kategorisk, oberoende variabel (responsvärde) med fler än två möjliga värden styrs av ett godtyckligt antal oberoende variabler (särdrag). Modellen byggs upp genom att räkna samförekomster av särdrag och responsvärden i träningsdata. Modellen antar att alla särdrag är oberoende av varandra. Vid klassificering av en ny observation väljs slumpmässigt ett av alla möjliga responsvärden ut som basvärde och utifrån observationens särdrag ges en sannolikhet att varje responsvärde skulle väljas istället för just detta basvärde. Det är en så kallad alla-mot-en-metod som gör antagandet att varje sannolikhetsfördelning mellan två responsvärden inte påverkas av vilka andra responsvärden som finns tillgängliga. Denna metod innebär att det bara krävs n - 1 uträkningar för att välja mellan n antal responsvärden, vilket kan spara väldigt mycket tid om det finns många responsvärden. En ingående beskrivning av multinomial logistisk regression ges av Jurafsky & Martin (2009, sida 235).

### 2.3.2 Namnlistor

Att bygga upp listor över olika typer av namn är en intuitiv lösning på namntaggningsuppgiften. Vid CoNLL-2003 användes en baslinjemetod som bestod av att endast använda träningsdatan för att bygga upp ett lexikon vilket fungerade relativt bra för uppgiften (Tjong & De Meulder, 2003).

Namnlistor är dock alltid begränsade i täckning och har svårt att hantera ambiguitet, men att använda matchningar i ordlistor som särdrag i en maskininlärningsalgoritm är fortfarande helt nödvändigt för att skapa bra namntaggningssytem (Florian et al., 2003).

### 2.3.3 Taggset

Vilka klasser som ska finnas med i ett system, eller ens vad som ska räknas som ett namn är aldrig helt självklart. MUC6 identifierade personer och institutioner (Grishman, 1996). CoNLL-2003 utökade med platser och en klass för övriga namn (Tjong & De Meulder, 2003). ACE definierade ett taggset med sju huvudklasser: person, organisation, plats, facilitet, vapen, fordon och geopolitisk entitet (Doddington et al., 2004). Dessutom delade de in alla klasserna i subklasser för en mer finkornig taggning. Klasserna vapen, facilitet och fordon kan verka något arbiträra. Orsaken till dessa är att det finns ett militärt intresse för namn- och entitetstaggning, och när namntaggningsystem skapas med specifika syften anpassas taggsetet efter syftet. Ett annat exempel är biomedicinska texter där man vill kunna tagga olika gener och proteiner som två olika typer av namn (Settles, 2005). I en mindre specifik modell kanske de inte skulle räknas som namn överhuvudtaget.

Eftersom namn kan bestå av flera token används ofta ytterligare ett lager information för att visa var ett namn börjar och var det slutar. Ett system för att lagra sådan information är BIO-schemat. Det första (eller enda) tokenet i ett namn taggas med ett B (beginning) och eventuellt följande token taggas med ett I (inside). Samtliga andra token taggas med O (outside). BIO-schemat är vanligt i namntaggning, men vid maskininlärning kan bättre namntaggning uppnås genom att utöka till BILOU-schemat, där L (last) markerar sista tokenet i ett namn, och U (unit-length) markerar namn bara består av ett token (Ratinov och Roth, 2009). Detta ökar dock antalet möjliga responsvärden, och har därför en viss negativ påverkan på maskininlärningens hastighet. Se tabell 1 för exempel på hur en och samma mening taggas på olika sätt med BIO- och BILOU-schema.

*Tabell 1. Exempelmening med två namn och jämförelse mellan BIO-schema och BILOU-schema. Obsververa att punkten är ett token som ingår i ett namn.*

| Token | BIO-schema | BILOU-schema |
|---|---|---|
| George | B | B |
| W | I | I |
| . | I | I |
| Bush | I | L |
| och | O | O |
| Obama | B | U |

## 2.4 Utvärdering av namntaggningssystem

### 2.4.1 Precision, täckning, och F1-värde

Precision och täckning är två nära besläktade mått som används för att beskriva hur väl en klassificeringsuppgift är utförd, vilket för namntaggning kräver att det finns en guldstandard att jämföra med. När ett system har taggat en text kontrolleras varje taggning för att kunna ta fram dessa mått. För varje klass k räknas först varje taggning t till en av följande fyra grupper:

- Sann positiv: taggning t med klass k finns i både guldstandard och taggning.
- Falsk positiv: taggning t med klass k finns i taggning men inte i guldstandard.
- Falsk negativ: taggning t med klass k finns i guldstandard men inte i taggning.
- Sann negativ: taggning t med klass k finns varken i guldstandard eller taggning.

Från dessa värden tas precision och täckning fram. Precision mäter andelen taggningar som är korrekta utav de taggningar som gjorts. Täckning mäter andelen korrekta taggningar som gjorts utav alla möjliga korrekta taggningar.

- Precision = Sann positiv / (Sann positiv + Falsk positiv)
- Täckning = Sann positiv / (Sann positiv + Falsk negativ)

Precision och täckning för ett helt system bör ske genom att först räkna ihop samtliga klasser och sedan ta fram deras totala precision och täckning. Att använda genomsnittet av de olika klassernas precision respektive täckning skulle inte ta hänsyn till att klasserna kan vara olika frekventa, och de mindre frekventa klasserna skulle då ha mer inflytande per taggning än de mer frekventa klasserna.

F1-värdet är det harmoniska medelvärdet av precision och täckning. Måttet är oförlåtande eftersom det alltid hamnar närmst det lägre av de två värdena, till skillnad från aritmetisk medelvärde som hamnar exakt mellan dem. På grund av detta vill man inte ha hög precision på bekostnad av täckning, eller vice versa. Det bästa resultatet uppnås genom att ha lika hög täckning och precision.

- F1 = 2 * (Precision * Täckning) / (Precision + Täckning)

### 2.4.2 K-faldig korsvalidering

Vid utveckling av övervakade system kan brist på uppmärkt data vara ett problem. Träningsdatan behöver vara stor nog för att bygga en bra modell och testdatan ska vara stor nog för att ge ett tillförlitligt resultat. Dessa två behov motverkar varandra. En lösning på problemet är *k-faldig korsvalidering,* där *k* antal utvärderingar genomförs. I varje utvärdering väljs en av *k* datamängder ut som testdata och resten används som träningsdata. Detta görs *k* gånger och varje datamängd används som testdata exakt en gång. Samtliga utvärderingar sammanställs sedan till ett slutgiltigt resultat, vilket kan ske med genomsnittet av utvärderingarnas resultat, eller genom att lägga ihop alla utvärderingar och ta fram det gemensamma resultatet. K-faldig korsvärdering används inte bara vid brist på data, utan också för att ta fram ett mer tillförlitligt och rättvist resultat (Stone, 1974).

### 2.4.3 Problematik med utvärdering

Det slutgiltiga betyget för ett namntaggningssystem är F1-värdet. Att ta fram precision och täckning för att kunna räkna ut F1-värdet vid namntaggning är dock inte fullt så trivialt som man kan tänka sig. I andra områden, som till exempel ordklasstaggning, låter man ett system tagga en text där det finns en manuellt annoterad guldstandard och jämför sedan varje taggat token med dess motsvarighet i guldstandarden. Denna metod är dock inte helt passande för utvärderingar av namntaggningssytem. Utvärdering sker fortfarande genom att jämföra den utförda taggningen med en guldstandard, men metoden måste modifieras av två anledningar.

För det första utgör namn en ganska liten andel av alla token i de flesta texter. Ett enkelt baslinjesystem som taggar samtliga token som ”inte namn” skulle ha både precision och täckning på cirka 90% om man räknar alla token. Av denna anledning brukar man endast räkna de token som är taggade som namn i antigen guldstandard eller utförd taggning. För det andra är en utvärdering på tokennivå problematiskt eftersom det man faktiskt letar efter är hela namn, och ett och samma namn kan ofta sträcka sig över flera token. På grund av detta utvärderas namntaggare inte på tokennivå, utan på namnnivå. Att sätta korrekt tagg på ett namn som bara består av ett token är alltså värt lika mycket som att sätta korrekt tagg på alla token i ett längre namn. Detta leder vidare till frågan hur man ska värdera delvis korrekta taggningar i namn med flera token. CoNLL-2003 löste detta den hårda vägen genom att endast räkna namn där alla token är korrekt taggade och både start- och sluttoken stämmer överens med guldstandard. (Tjong & De Meulder, 2003).

## 2.5 Stockholm-Umeå Corpus

Stockholm-Umeå Corpus (Källgren, 2006), förkortat SUC, är en balanserad svenskspråkig textkorpus. Den är baserad på Brown-modellen (Francis & Kučera, 1964) och innehåller cirka en miljon token i 500 texter från nio väldigt olika domäner: reportage, ledare, recensioner, yrkes- och hobbymagasin, populärvetenskap, biografier och essäer, facklitteratur, skönlitteratur, och misc (regeringspublikationer, kommunpublikationer, finansrapporter, interna företagspublikationer, och universitetspublikationer). De olika domänerna innehåller inte lika många texter. Den minsta domänen (ledare) innehåller endast 17 texter, medan den största domänen (skönlitteratur) innehåller 127 texter. Även inom domänerna är texterna varierade. Se bilaga A för hela listan av SUC-domäner.

Alla texter i SUC är ordklasstaggade och kontrollerade manuellt. Den senaste versionen av korpusens annotering - SUC 3.0 (Östling, 2013) - innehåller över 1000 rättade felaktiga ordklasstaggningar som hittats i den tidigare utgåvan. Korpusen har dessutom en manuellt utförd namntaggning som består av klasserna person, djur, plats, institution, verk, produkt, händelse, myt, och övrigt. I denna namntaggning tillåts samma namn tillhöra olika klasser i olika kontexter. (Källgren, 2006). SUC 3.0 är tillgänglig i bland annat ett CoNLL-X-baserat format[1], och använder BIO-schemat för att markera namngränser. SUC är den enda fritt tillgängliga svenska korpusen med manuellt utförd namntaggning.

[1] *http://ilk.uvt.nl/conll/#dataformat, 15 maj 2016*

## 2.6 Namntaggningssystem för svenska

Flera system har tagits fram för namntaggning av svensk text, med varierande taggset. Kokkinakis (1998) beskrev ett regelbaserat system som identifierar personer, platser, organisationer, samt ett antal klasser speciellt definierade för texter som behandlar narkotikarelaterad brottslighet. Dalianis & Åström (2001) använde både maskininlärning och handskrivna regler för att hitta personer, platser, organisationer och tidsuttryck. Ek et al. (2011) beskrev ett system som är speciellt utvecklat för sms-meddelanden som identifierar personnamn, platser, tidsuttryck, datum, och telefonnummer.

Bland de nyare systemen finns det åtminstone två som är jämförbara i taggset och utverderingsdata med det system som tagits fram för denna studie.

Salomonsson et al. (2012) tog fram ett system som fick ett F1-värde på 74,0% vid tiofaldig korsvalidering på SUC med samma taggset och utvärderingsmetod som CoNLL-2003. Detta innebär att alla namnklasser förutom person, plats, och institution slogs ihop till en stor övrigt-klass. De lät dessutom institution heta organisation, utan att förändra innehållet. Systemet de utvecklade finns inte tillgängligt för användning[2], och hela systemet finns inte beskrivet, men utvärderingen finns publicerad.

Östling (2013) inkluderade funktionalitet för namntaggning i ordklasstaggaren Stagger. Han använde samma taggsetsmetod och utvärderingsstandard som Salomonsson et al. (2012) och fick ett F1-värde på 70,7%.

[2] *Salomonsson, personlig kommunikation, 14 mars 2016*

# 3 Syfte och frågeställningar

## 3.1 Syfte

Syftet med uppsatsen är att utveckla en övervakad namntaggare för att kunna undersöka om den uppnår bättre resultat genom att utesluta träningsdata som inte tillhör samma textdomän som testdatan.

## 3.2 Frågeställningar

Frågeställning 1: Hur påverkas precision, täckning och F1-värde för det framtagna systemets namntaggning av att en mindre mängd endast domänspecifik träningsdata från SUC används vid multinomial logistisk regression?

Hypotes 1: Användning av endast domänspecifik träningsdata ger bättre precision, täckning och F1-värde.

Frågeställning 2: Påverkas alla namnklasser likadant (i samma riktning)?

Hypotes 2: Alla namnklasser påverkas likadant.

Frågeställning 3: Påverkas alla textdomäner likadant (i samma riktning)?

Hypotes 3: Alla textdomäner påverkas likadant.

# 4 Metod

## 4.1 Systemarkitektur

Namntaggningssystemet bygger på en multinomial logistisk regressionmodell som implementeras i Python 3.4.2 med biblioteket SciKit-Learn 0.17.1 (Pedregosa, 2011). Inga parametrar ändras.

### 4.1.1 Träningsdata och särdrag

Systemet tar emot träningsdata i det CoNLL-X-baserade format som SUC 3.0 (Östling, 2013) använder. Den information som faktiskt används av systemet är ytform, ordklasstaggning, meningsintern tokenindexering, och själva namntaggningen. Vissa modifikationer utförs på datan. Det namntaggset som finns i SUC reduceras till de fyra klasserna person, plats, institution och övrigt, i likhet med Salomonsson et al. (2012) och Östling (2013). BIO-schemat utökas till BILOU genom att alla B-taggar som inte följs av I skrivs om till U och alla I-taggar som inte följs av I skrivs om till L. För varje observerat token i träningsdatan konkateneras BILOU-taggen med namnklassen. Denna konkatenering används som responsvärde för observationen. De särdrag som kopplas till varje observation är:

- Observerat tokens ytform.
- Konkatenering av föregående och observerat tokens ytformer.
- Konkatenering av observerat och följande tokens ytformer.
- Observerat tokens ordklass.
- Konkatenering av föregående och följande tokens ordklasser.
- Föregående tokens responsvärde.
- Observerat tokens versaliseringsmönster.
- Följande tokens versaliseringsmönster.
- Observerat token finns i personnamnlista.
- Observerat token finns i platsnamnlista.

Särdragen som behandlar versaliseringsmönster kan ta ett av sju värden: inga versaler (xxx), endast versaler (XXX), versaliserad (Xxx), meningsinitialt versaliserad (. Xxx), versaliserad efter citattecken (” Xxx), versaliserad efter bindestreck (- Xxx), eller okänt versaliseringsmönster (xXx,, xxX, xXX). De särdrag som jämför token med namnlistor kan ta två värden: sant eller falskt. Värdet sant anges vid exakt matchning (Malmö = Malmö), matching efter borttaget s i slutet av ytformen (Malmös = Malmö), eller matching efter att ett token med endast versaler ges normal versalisering (MALMÖ = Malmö). Ett vanligt namn utan versalisering, till exempel ”joakim”, skulle däremot inte matchas.

### 4.1.2 Framtagna namnlistor

Två ordlistor har tagits fram för systemet: personnamn och platsnamn. Personnamnlistan består av de 600 vanligaste svenska pojknamnen, de 600 vanligaste svenska flicknamnen, och de 600 vanligaste svenska efternamnen, enligt Allt För Föräldrar[3]. Med vanligaste namnen avses inte de mest frekventa namnen för nyfödda barn, utan de mest frekventa namnen för hela populationen i Sverige. Platsnamnlistan består av alla svenska tätorter och kommuner[4], landskap och landsdelar; alla världsdelar, länder och huvudstäder[5]; samt alla amerikanska delstater. Detta utgör cirka 2400 platsnamn.

### 4.1.3 Sekventiell namntaggning

Den text som ska namntaggas måste precis som träningsdatan vara tokeniserad, ordklasstaggad, och ha meningsintern tokenindexering för att särdrag ska kunna extraheras och token ska kunna taggas. Vid taggning av text används en enkel sekventiell metod. Varje token taggas en gång, och det givna responsvärdet blir även indata till nästa observation, eftersom ett av särdragen är föregående observations responsvärde. Inga taggningar förändras i efterhand. När alla token är taggade stegas de igenom igen för att ta fram mängden av alla namn. Om ett token har BILOU-tagg U så utgör det ett namn på egen hand. Om ett token har BILOU-tagg B så utgör det början på ett namn som fortsätter så länge följande token tillhör samma klass och har BILOU-tagg I eller L (vid O eller efter L måste ett nytt B eller U förekomma innan ett nytt namn räknas). Detta innebär att systemet ignorerar eventuella taggningar som inleds med I eller L.

## 4.2 Utvärdering av systemet

### 4.2.1 Jämförelse av namnmängder

Efter att mängden av namn har tagits fram i både i den gjorda taggningen och i guldstandarden jämförs dessa mängder. De namn som förekommer i identisk form i båda mängderna räknas som fullständigt korrekt taggade namn och tas bort från båda mängderna. För att sedan räkna delvis korrekt taggade namn jämförs de återstående namnen från guldstandarden med de återstående namnen från taggningen. Om det förekommer åtminstone ett gemensamt token med samma klass i båda mängderna så räknas det som en delvis korrekt taggning och båda namnen tas bort från mängderna. Ett och samma namn kan alltså endast utgöra en enda delvis korrekt matchning, även om det överlappar med flera. Om ett namn som består av flera token blir felaktigt taggat som två namn av systemet räknas det alltså som en delträff och en feltaggning. Om två namn som följer på varandra blir felaktigt taggade som ett namn av systemet räknas det som en delträff och ett missat namn.

Eftersom utvärderingen sker på två nivåer av noggrannhet (fullständigt och delvis korrekt) innebär det att två slutgiltiga utvärderingsmått tas fram. Det första måttet räknar endast fullständiga träffar, enligt samma krav som CoNLL-2003. Det andra måttet räknar både fullständigt korrekt taggade namn och delvis korrekt taggade namn.

---

[3] http://svenskanamn.alltforforaldrar.se/statistik/statistics/ *(7 april 2016)*

[4] http://scb.se/Statistik/MI/MI0810/2010A01T/MI0810_To_So_Kommun2010.xls *(29 mars 2016)*

[5] https://sv.wikipedia.org/wiki/Lista_över_huvudstäder_efter_land *(29 mars 2016)*

### 4.2.2 Korsvalidering

Namntaggningssystemet utvärderas med en tiofaldig korsvalidering över alla 500 texter i SUC. I varje körning består testdatan av 50 texter och träningsdatan består av resterande 450 texter. Urvalet sker genom att texterna sorteras alfanumeriskt och var tionde text används som testdata. Inför varje ny körning förskjuts urvalet ett steg. På så sätt används samtliga 500 texter som testdata vid någon av körningarna. Användning av var tionde text (istället för slumpmässigt urval) ger så balanserat resultat som möjligt (Sjöbergh, 2003). Resultaten från alla körningar summeras och ett stort resultat räknas ut från dessa summor.

## 4.3 Experimentdesign

Den experimentella delen av studien består av två variationer av en och samma namntaggningsuppgift. Denna uppgiften innehåller nio block av tiofaldiga korsvalideringar som görs enligt samma urvalsmetod som beskrivits ovan i sektion 4.2.2. Varje block innehåller testdata från endast en textdomän i SUC. Inga block innehåller överlappande textdata. De två variationerna utförs med exakt samma system som beskrivits ovan i sektion 4.1.

Båda varianterna har gemensamt att de använder den träningsdata som finns tillgänglig inom samma domän som testdatan, men det som skiljer dem åt är att den ena varianten *endast* använder denna träningsdata, medan den andra varianten dessutom använder samtliga textfiler från de övriga åtta kategorierna i SUC. De två körningarna utförs med en 2,4 GHz Intel Xeon E5645-processor. Körtider mäts med Unix-kommandot *time* och avrundas till minuter.

Resultat presenteras med F1-värden. Signifikanstester utförs på precision respektive täckning (se avsnitt 2.4.1) med de absoluta värden som uppmäts i körningarna. Både precision och täckning jämförs inom både fullständigt korrekt taggande namn och åtminstone delvis korrekt taggade namn. Testet som används är Pearsons chi-2-test. De två oberoende grupperna är domänspecifik taggning och icke domänspecifik taggning i samtliga fyra signifikanstester. De ömsesidigt uteslutande mätvariablerna för precision är antalet sanna positiver respektive antalet falska positiver, vilket summeras till det totala antalet taggade namn. Motsvarande mätvariabler för täckning är antal sanna positiver respektive antalet falska negativer, vilket summeras till antalet namn i guldstandarden.

# 5 Resultat

## 5.1 Experimentresultat

### 5.1.1 Resultat utifrån frågeställningar

1: Användningen av domänspecifik träningsdata gav i experimentet en försämring på 1,9 procentenheter i F1-värde i den totala sammanräkningen för fullständigt korrekt taggade namn (se tabell 1), och en försämring på 1,3 procentenheter i F1-värde i den totala sammanräkningen för åtminstone delvis korrekt taggade namn. Både precision och täckning visade en signifikant försämring med $p < 0,01$ för både fullständigt korrekt taggade namn och åtminstone delvis korrekt taggade namn (se bilaga B).

2: Två av de fyra namnklasserna (person och institution) fick förbättrade resultat av domänspecifik träningsdata, medan de två andra (plats och övrigt) fick sämre resultat (se tabell 1 och 2).

3: Två av de nio textdomänerna – ledare och misc - fick förbättrade resultat för fullständigt korrekt taggade namn (se tabell 1), varav misc även fick förbättrat resultat för åtminstone delvis taggade namn (se tabell 2).

Inga av de tre framlagda hypoteserna kunde styrkas av experimentresultaten. Att träna, använda och utvärdera systemet med domänspecifik träningsdata för nio stycken tiofaldiga korsvalideringar tog 41 minuter. Körningen med ej domänspecifik träningsdata tog 11 timmar och 16 minuter.

### 5.1.2 Fullständigt korrekt taggade namn

Den icke domänspecifika taggningen fick F1-värdet 67,6% på fullständigt korrekt taggade namn, och slog därmed den domänspecifika taggningen med 1,9 procentenheter (se tabell 1).

*Tabell 1. Båda körningarnas F1-värden för fullständigt korrekt taggade namn. I varje cell anges resultat för domänspecifik träningsdata till vänster och resultatet för icke domänspecifik träningsdata till höger. I de grå cellerna har domänspecifik träningsdata gett högre F1-värde än icke domänspecifik träningsdata.*

| | Person | Plats | Institution | Övrigt | TOTALT |
|---|---|---|---|---|---|
| Reportage | 79,3% - 78,3% | 72,8% - 74,8% | 41,4% - 41,5% | 12,6% - 20,0% | 64,8% - 67,1% |
| Ledare | 71,4% - 70,4% | 64,3% - 70,0% | 66,1% - 57,6% | 0,0% - 25,6% | 65,3% - 64,3% |
| Recensioner | 81,4% - 82,5% | 61,9% - 64,1% | 5,9% - 23,3% | 11,5% - 17,3% | 66,7% - 67,5% |
| Yrke & hobby | 75,2% - 69,7% | 60,6% - 68,0% | 46,0% - 46,7% | 13,2% - 20,8% | 56,3% - 59,9% |
| Popvetenskap | 70,3% - 78,5% | 74,6% - 78,0% | 14,8% - 32,3% | 31,4% - 34,3% | 64,0% - 69,5% |
| Biografi & essä | 78,1% - 84,3% | 68,3% - 69,8% | 0,0% - 34,3% | 8,0% - 25,0% | 67,9% - 72,4% |
| Misc | 75,2% - 52,3% | 78,6% - 81,4% | 41,9% - 38,2% | 37,9% - 40,3% | 63,2% - 60,8% |
| Vetenskap | 61,5% - 67,4% | 67,3% - 69,6% | 10,1% - 27,0% | 17,0% - 23,7% | 52,6% - 57,8% |
| Skönlitteratur | 68,1% - 86,6% | 74,4% - 74,7% | 0,0% - 16,5% | 54,0% - 52,9% | 80,3% - 80,9% |
| TOTALT | 78,5% - 78,3% | 71,3% - 73,9% | 39,4% - 39,0% | 23,7% - 29,4% | 65,7% - 67,6% |

Personklassen förbättrades av domänspecifik träningsdata inom fyra av nio domäner, och även i den totala sammanräkningen. Platsklassen försämrades i samtliga kategorier. Institutionsklassen förbättrades som helhet trots att den försämrades inom sju av nio kategorier. Övrigt-klassen försämrades i alla domäner förutom skönlitteratur. Textdomänerna ledare och misc var de enda som hade helhetsförbättring, och gemensamt hade de att de förbättrades i både person- och institutionsklassen (se tabell 1).

### 5.1.3 Åtminstone delvis korrekt taggade namn

Även för åtminstone delvis korrekt taggade namn gav icke domänspecifik träningsdata ett bättre resultat, men då endast med 1,3 procentenheter (se tabell 2).

*Tabell 2. Båda körningarnas F1-värden för åtminstone delvis korrekt taggade namn. I varje cell anges resultat för domänspecifik träningsdata till vänster och resultatet för icke domänspecifik träningsdata till höger. I de grå cellerna har domänspecifik träningsdata gett högre F1-värde än icke domänspecifik träningsdata.*

| | Person | Plats | Institution | Övrigt | TOTALT |
|---|---|---|---|---|---|
| Reportage | 84,4% - 83,3% | 74,3% - 76,8% | 49,3% - 48,5% | 18,5% - 26,9% | 69,5% - 71,6% |
| Ledare | 82,0% - 82,8% | 64,3% - 71,4% | 70,8% - 61,9% | 2,9% - 30,2% | 70,8% - 71,0% |
| Recensioner | 86,4% - 89,1% | 63,1% - 66,0% | 7,8% - 30,8% | 25,2% - 27,5% | 72,1% - 73,8% |
| Yrke & hobby | 81,9% - 74,3% | 66,5% - 70,1% | 49,9% - 50,9% | 19,2% - 27,3% | 61,9% - 63,9% |
| Popvetenskap | 75,5% - 82,9% | 77,9% - 81,4% | 16,8% - 40,2% | 36,3% - 41,9% | 68,1% - 74,0% |
| Biografi & essä | 84,5% - 86,5% | 69,0% - 72,8% | 0,0% - 38,7% | 9,6% - 28,9% | 72,8% - 75,1% |
| Misc | 85,6% - 56,1% | 80,3% - 83,0% | 55,3% - 48,2% | 37,9% - 44,0% | 69,7% - 65,1% |
| Vetenskap | 68,7% - 72,0% | 67,7% - 70,3% | 12,6% - 35,7% | 23,5% - 29,8% | 57,4% - 62,0% |
| Skönlitteratur | 91,1% - 91,9% | 76,2% - 76,6% | 0% - 19,9% | 55,8% - 57,7% | 84,5% - 84,8% |
| TOTALT | 84,3% - 83,1% | 73,4% - 75,8% | 46,2% - 45,8% | 29,3% - 35,7% | 70,7% - 72,0% |

Både personklassen och institutionsklassen förbättrades som helhet, men endast inom tre av nio domäner. Platsklassen och övrigt-klassen försämrades i samtliga domäner. Misc är den enda textdomänen som förbättrades totalt (se tabell 2).

## 5.2 Systemets utvärderingsresultat

Den balanserade tiofaldiga korsvalidering på SUC som genomfördes med systemet gav ett F1-värde på 67,5% för fullständigt korrekt taggade namn och 71,7% för åtminstone delvis korrekt taggade namn (se tabell 3).

*Tabell 3. Precision, täckning och F1-värde för de olika namnklasserna. Det första måttet mäter endast fullständigt korrekt taggade namn, det andra mäter åtminsone delvis korrekt taggade namn.*

| | Person | Plats | Institution | Övrigt | TOTALT |
|---|---|---|---|---|---|
| Precision | 70,7% / 74,9% | 69,5% / 71,3% | 59,9% / 70,0% | 50,5% / 61,7% | 68,5% / 72,7% |
| Täckning | 87,7% / 92,8% | 78,6% / 80,6% | 28,5% / 33,3% | 19,7% / 24,0% | 66,5% / 70,7% |
| F1-värde | 78,3% / 82,9% | 73,8% / 75,6% | 38,6% / 45,1% | 28,3% / 70,7% | 67,5% / 71,7% |

Person och plats är de två mest frekventa klasserna i SUC. Tillsammans utgör de 70% av namnen. Systemet har överskattat denna snedfördelning och satt antingen person- eller platstagg på 86% av alla hittade namn (se tabell 4).

*Tabell 4. Absoluta frekvenser för de olika namnklasserna från den tiofaldiga korsvalideringen.*

| | Person | Plats | Institution | Övrigt | TOTALT |
|---|---|---|---|---|---|
| Antal | 15128 | 8775 | 6334 | 3955 | 34192 |
| Hittade | 18761 | 9918 | 3017 | 1539 | 33235 |
| Helt korrekt | 13272 | 6896 | 1806 | 777 | 22751 |
| Delvis korrekt | 771 | 173 | 305 | 173 | 1422 |

# 6 Diskussion

## 6.1 Metoddiskussion

### 6.1.1 Diskussion om maskininlärningsmetod

Multinomial logistisk regression är en etablerad metod för kategoriska klassificeringsuppgifter inom datorlingvistik (Jurasky & Martin, 2009, sida 235). Den använder sig av ett pseudoslumpmässigt genererat nummer som kan leda till små skillnader i sannolikhetsuträkningar vid olika körningar med samma träningsdata. Detta är dock av sådan liten grad att det inte bör ha någon negativ effekt på studiens reliabilitet.

Inga av de särdrag som används i systemet är ovanliga eller innovativa. De resultat som uppnåtts vid systemets utvärdering (se avsnitt 5.2) är i den undre närheten av liknande namntaggningsytems (se avsnitt 2.6), vilket gör det rimligt att tro att även de skulle kunna få liknande resultat ifall samma experiment utfördes med dem och med samma träningsdata.

### 6.1.2 Diskussion om Stockholm-Umeå Corpus

Stockholm-Umeå Corpus innehåller texter av väldigt olika karaktär. Det är ett medvetet val som gjorts för att SUC ska representera svenska språket på ett så balanserat och brett sätt som möjligt. Detta har sina för- och nackdelar när den används som träningsdata. Till exempel kan skillnaderna mellan domänerna vara så pass stora att den eventuella negativa effekten som blandade domäner har på maskininlärning förstärks och överdrivs. Den viktiga frågan är här ifall det någonsin är rimligt att träna en namntaggare på en så balanserad textsamling som SUC? Kan man ta med sig resultatet från detta experiment på SUC och anse att det representerar verkligheten? Detta kan eventuellt ses som ett hot mot validiteten i studien.

Vissa inkonsekvenser av namntaggning har upptäckts i SUC. Titlar som herr och fru ingår ibland i namn, och ibland inte (Salomonsson et al., 2012). Inkonsekvenser som detta kan vara förödande för en namntaggare som letar efter mönster i träningsdata. Ett antal direkta feltaggning har också upptäckts. Dessa sorts fel är dock mer eller mindre oundvikliga för en korpus och ger i bästa fall bara lite mer brus i resultatet, så länge feltaggningar av samma sort inte upprepar sig systematiskt.

### 6.1.3 Diskussion om använt taggset

Att använda taggsetet person, plats, institution och övrigt har blivit mer eller mindre standard. Större delen av alla namn som dyker upp i texter kan påstås tillhöra någon av de tre första klasserna. Det finns dock ett relativt stort problem med detta taggset, åtminstone som det är tillämpat i SUC, och det är den inkonsekventa taggningen av länder. Anledningen till detta är distinktionen mellan ett land som geografisk plats och politisk aktör. Det är inte konstigt att man vill kunna göra denna distinktion, för ibland är den ena eller den andra taggen helt opassande i kontexten. Till exempel skulle det vara opassande att tagga ”Frankrikes närvaro i Afrika” med två platstaggar eftersom Frankrike här inte syftar på platsen Frankrike.

Problemet som uppstår är att systemet ser samma sträng taggas många gånger med två olika taggar. Det blir sedan nästan godtyckligt ifall landsnamn kommer taggas som plats eller institution. Dessutom kan systemet bli väldigt felträ nat ifall det kopplar ihop för många institutionstaggar med namn som finns med på platsnamnlistan. En lösning på detta är att använda en separat tagg för namn som kan vara både plats och institution, vilket är precis vad man gör i ACE (se avsnitt 2.2) med taggen geopolitisk entitet.

### 6.1.4 Diskussion om använda namnlistor

De två namnlistor som används i systemet är förhållandevis små (1800 personnamn och 2400 platsnamn), vilket ger låg täckning men hög precision. Endast 23 strängar finns med på båda listorna: Berg, Berlin, Dorotea, Fors, Fredrika, Göta, Hammar, Holm, Hult, Ljung, Lund, Mark, Nora, Norberg, Nordmark, Sjöberg, Skog, Sofia, Sund, Ulrika, Victoria, Vilhelmina, och Åsa.

Vid sammanställningen av listorna visade det sig att överlapp började ske i mycket större utsträckning efter ungefär 600 namn från personnamnlistorna, vilket ledde till beslutet att inte göra listorna längre. Listor som innehåller fler namn ger bättre täckning men sämre precision. En tänkbar framtida lösning på detta är att skapa flera listor för samma namnklass, där en mindre lista innehåller de mest frekventa namnen och har mycket hög precision, medan en annan lista innehåller väldigt många namn och därmed har hög täckning.

Platsnamnlistan har en del uppenbara brister. För det första innehåller den inga icke-svenska städer som inte är huvudstäder. Detta skulle enkelt kunna förbättras, men listan skapades med syftet att bara täcka de mest självklara namnen, och gränsen var tvungen att dras någonstans. Eftersom listan dessutom är avsedd att vara så grundläggande som möjligt så är det en del mycket enkla namn från SUC som inte täcks av listan. SUC gavs ut på 1990-talet, och forna länder som Sovjet och Tjeckoslovakien finns inte med på platslistan, eftersom den endast innehåller de länder som finns idag.

### 6.1.5 Diskussion om experimentets utformning

Syftet med studien var att ta reda på om bättre namntaggning kunde uppnås genom att ta bort all träningsdata som inte tillhör samma domän som testdatan. Att utföra separata körningar för varje domän gav möjligheten att se ifall alla domäner påverkades olika, vilket var helt nödvändigt för att kunna besvara frågeställning 3.

Dock togs ingen hänsyn till domänernas storlek. Antalet textdokument som ingick i mängden domänspecifik träningsdata var nästan åtta gånger större för skönlitteratur än för ledare, den största respektive minsta domänen. Det hade varit bra att normalisera träningsdatans storlek, och helst utföra tester med flera olika storlekar för att se ifall det finns några kritiska storleksmängder där den positiva effekten av mer träningsdata avtar, både för domänspecifik, icke domänspecifik och uteslutande ej domänspecifik träningsdata.

## 6.2 Resultatdiskussion

### 6.2.1 Diskussion om huvudresultat

Resultaten från experimentet visar på en statistiskt signifikant försämring av namntaggning, både för fullständigt korrekt taggade och åtminstone delvis taggade namn, när endast den domänspecifika träningsdatan används. Skillnaden är dock relativt liten (mindre än två procentenheter i båda fallen) och dess egentliga värde kan ifrågasättas med tanke på körtiderna, som visserligen inte ingick i frågeställningen, men ändå är värda att ta upp med tanke på hur liten skillnaden är mellan körningarnas resultat.

Den icke domänspecifika körningen tog nämligen 16 gånger längre tid att genomföra än den domänspecifika, men dess F1-värde var endast 1,029 gånger bättre på fullt korrekt taggade namn och 1,018 gånger bättre på åtminstone delvis korrekt taggade namn, och i vissa namnklasser och textdomäner blev F1-värdet till och med bättre med domänspecifik träningsdata. Om man har för avsikt att träna sitt system upprepade gånger och tid är en begränsad resurs bör man fundera på om den extra körtiden är värd förbättringen eller inte. I vissa fall kanske det inte bara är tiden som är begränsad, utan även tillgången till manuellt annoterad data. Att ta fram ny träningsdata tar lång tid och kostar mycket pengar eftersom det krävs manuellt arbete.

### 6.2.2 Diskussion om resultat utifrån namnklasser

Att två av de fyra namnklasserna (person och institution) förbättrades tyder på att man faktiskt kan uppnå både snabbare och bättre namntaggning genom att ta bort överflödig, irrelevant träningsdata, åtminstone i viss utsträckning. Detta rimmar bra med de goda resultat som uppmättes på MUC6 (Grishman & Sundheim, 1996), när både träningsdatan och testdatan bestod av Wall Street Journal-texter, och endast personer och institutioner skulle taggas. Persontaggen är kanske den viktigaste av alla taggar, vilket återspeglas i att nästan hälften av alla taggade namn i SUC är personnamn. Person och institution har förbättras totalt, men i färre än hälften av domänerna. Det är alltså i de domäner där de är som mest frekventa som de har förbättras. Detta skulle kunna bero på att de fortfarande har en tillräcklig mängd träningsdata där, även med domänspecifik träningsdata. Samma förklaring skulle gå att applicera på övrigt-klassen, som försämrades rakt igenom. Även detta kan eventuellt bero på att den är mindre frekvent än de andra klasserna, och att den helt enkelt får för få observationer med domänspecifik träningsdata. Detta förklarar dock inte varför plats försämras i alla domäner, eftersom det är den näst största klassen. Möjligtvis är plats den klass som är mest homogen över olika domäner, vilket faktiskt innebär att mer data alltid är bättre data i det fallet.

### 6.2.3 Diskussion om resultat utifrån textdomäner

Eftersom ledare är den minsta av alla textdomäner i SUC är det något överraskande att den var en av de två domäner som fick bättre resultat av domänspecifik träningsdata. En möjlig förklaring till detta skulle kunna vara att texterna är mer lika varandra inom domänen ledare än texter inom andra domäner. Frågan är dock om man kan säga detsamma om den andra domänen som fick förbättrade resultat: misc (regeringspublikationer, kommunpublikationer, finansrapporter, interna företagspublikationer, och universitetspublikationer). Denna domän är för övrigt den tredje största, så storlek tycks inte vara någon ledtråd till varför dessa två var de enda som förbättrades.

# 7 Slutsatser

Syftet med denna uppsats har varit att undersöka ifall övervakad namntaggning kan göras bättre genom uteslutning av träningsdata som inte tillhör samma textdomän som testdatan. För att genomföra detta har en namntaggare som använder multinomial logistisk regression utvecklats och utvärderats med en tiofaldig korsvalidering på Stockholm-Umeå Corpus. Tre frågeställningar formulerades angående effekten av domänspecifik träningsdata. Utav de tre hypoteserna som gavs med frågeställningarna kunde ingen styrkas av studien.

Den domänspecifika träningsdatan gav ett lägre F1-värde för både fullständigt och åtminstone delvis korrekt taggade namn. Skillnaden var signifikant för både täckning och precision, men inte större än två procentenheter för F1-värdet. Denna försämring bör dock ställas i proportion mot den 16-faldiga förbättring i körtid som den domänspecifika träningsdatan gav.

Förutom att försämringen är liten så är den heller inte genomgående. Person- och institutionsnamn fick bättre F1-värde av domänspecifik träningsdata, medan klasserna plats och övrigt försämrades. Av de nio textdomänerna som finns i SUC förbättrades F1-värdet för ledare och misc i mätningarna för fullständigt korrekt taggade namn.

För vidare forskning av liknande sort bör träningsmängder normaliseras i storlek och många olika storlekar av träningsdata bör användas för att se ifall det finns en kritisk punkt där mer data inte längre ger mycket bättre resultat, både för domänspecifik, icke domänspecifik, och uteslutande icke domänspecifik träningsdata. För att sedan kunna utnyttja domänspecifik träningsdata på okända texter skulle namntaggning kunna föregås av en textklassificering som avgör vilken del av träningsdatan som är mest lämplig att träna systemet med. I praktiken innebär detta att flera olika system är tränade i förväg, och att klassificeringen helt enkelt avgör vilket system som ska användas. En textklassificering kommer dock att motverka den positiva tidseffekten som man tjänar på att använda domänspecifik träningsdata, men förhoppningsvis inte mer än att det ändå är värt det.

# Referenser


Dalianis, H. & Åström, E. (2001). SweNam – a Swedish named entity recognizer. Technical Report. Department of Numerical Analysis and Computing Science, TRITA-NA-P0113 – IPLab-189. Stockholm, Sweden. <ftp.nada.kth.se/IPLab/TechReports/IPLab-189.pdf>

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., & Weischedel, R. M. (2004). The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. *LREC, 2*, 1-4.

Ek, T., Kirkegaard, C., Jonsson, H., & Nugues, P. (2011). Named entity recognition for short text messages. *Procedia-Social and Behavioral Sciences*, *27*, 178-187.

Elson, D. K., Dames, N., & McKeown, K. R. (2010). Extracting social networks from literary fiction. *Proceedings of the 48th annual meeting of the association for computational linguistics,* 138-147. Association for Computational Linguistics.

Florian, R., Ittycheriah, A., Jing, H. & Zhang, T. (2003). Named entity recognition through classifier combination. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, 4*, 168-171. Association for Computational Linguistics.

Francis, W. N. & H. Kučera. (1964). Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Providence, Rhode Island: Department of Linguistics, Brown University. Revised 1971. Revised and amplified 1979.

Grishman, R. & Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. *COLING, 96*, 466-471.

Jurafsky, D. & Martin, JH. (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. New Jersey: Prentice-Hall.

Kokkinakis, D. (1998). AVENTINUS, GATE and Swedish Lingware. *Proceedings of NoDaLiDa*.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Frontiers in Artificial Intelligence and Applications, 160*, 3-24.

Källgren, G. (2006). Documentation of the Stockholm-Umeå Corpus. *Manual of the Stockholm Umeå Corpus version 2.0.* Sofia Gustafson-Capková and Britt Hartmann (red). Stockholm University: Department of Linguistics.

Lee, D. Y. (2001). Genres, Registers, Text Types, Domain, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning & Technology*, *5*(3), 37-72.

Tjong Kim Sang, E. F. & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language- independent named entity recognition. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003,* 4, 142-147. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M. &

Perrot, M. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830.

Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning,* 147-155. Association for Computational Linguistics.

Settles, B. (2005). ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, *21*(14), 3191-3192.

Sjöbergh, J. (2003). Combining POS-taggers for improved accuracy on Swedish text. *Proceedings of NoDaLiDa, 2003*.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, 111-147.

Östling, R. (2013). Stagger: An open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology (NEJLT)*, *3*, 1-18.

# Bilagor

## A: SUC-domäner

Domäner enligt originalnamn, antalet texter per huvuddomän, och subdomäner enligt originalnamn (Källgren, 2006).

**A - Press: Reportage (44)**

AA Political. AB Community. AC Financial. AD Cultural. AE Sports. AF Spot News

**B - Press: Editorial (17)**

BA Institutional. BB Debate articles

**C - Press: Reviews (27)**

CA Books. CB Films. CC Art. CD Theatre. CE Music. CF Artists, shows. CG Radio, TV

**E - Skills and Hobbies (58)**

EA Hobbies, amusements. EB Society press. EC Occupational and trade union press. ED Religion

**F - Popular Lore (48)**

FA Humanities. FB Behavioural sciences. FC Social sciences. FD Religion. FE Complementary life styles. FF History. FG Health and medicine. FH Natural science, technology. FJ Politics. FK Culture

**G - Biographies, essays (26)**

GA Biographies, memoirs. GB Essays

**H - Miscellaneous (70)**

HA Government publications. HB Municipal publications. HC Financial reports, business. HD Financial reports, non-profit organisations. HE Internal publications, companies. HF University publications

**J - Learned and scientific writing (83)**

JA Humanities. JB Behavioural sciences. JC Social sciences. JD Religion  JE Technology. JF Mathematics. JG Medicine. JH Natural science, technology.

**K - Imaginative prose (127)**

KK General fiction. KL Science fiction and mystery. KN Light reading. KR Humour

# B: Signifikanstester

Absoluta frekvenser från experimentet. Denna data används till Pearsons chi-2-test.

*Tabell 5. Precision för fullständigt korrekt taggade namn.*

| | Domänspecifik | Icke domänspecifik |
|---|---|---|
| Sann positiv | 22116 | 22765 |
| Falsk positiv | 11026 | 10410 |

*Tabell 6. Täckning för fullständigt korrekt taggade namn.*

| | Domänspecifik | Icke domänspecifik |
|---|---|---|
| Sann positiv | 22116 | 22765 |
| Falsk negativ | 12076 | 11427 |

*Tabell 7. Precision för åtminstone delvis korrekt taggade namn.*

| | Domänspecifik | Icke domänspecifik |
|---|---|---|
| Sann positiv | 23796 | 24249 |
| Falsk positiv | 9346 | 8926 |

*Tabell 8. Täckning för åtminstone delvis korrekt taggade namn.*

| | Domänspecifik | Icke domänspecifik |
|---|---|---|
| Sann positiv | 23796 | 24249 |
| Falsk negativ | 10396 | 9943 |

Stockholms universitet
SE-106 91 Stockholm
Telefon: 08 – 16 20 00
www.su.se

Stockholms universitet