



Pre-editering och maskinöversättning

Convertus AB

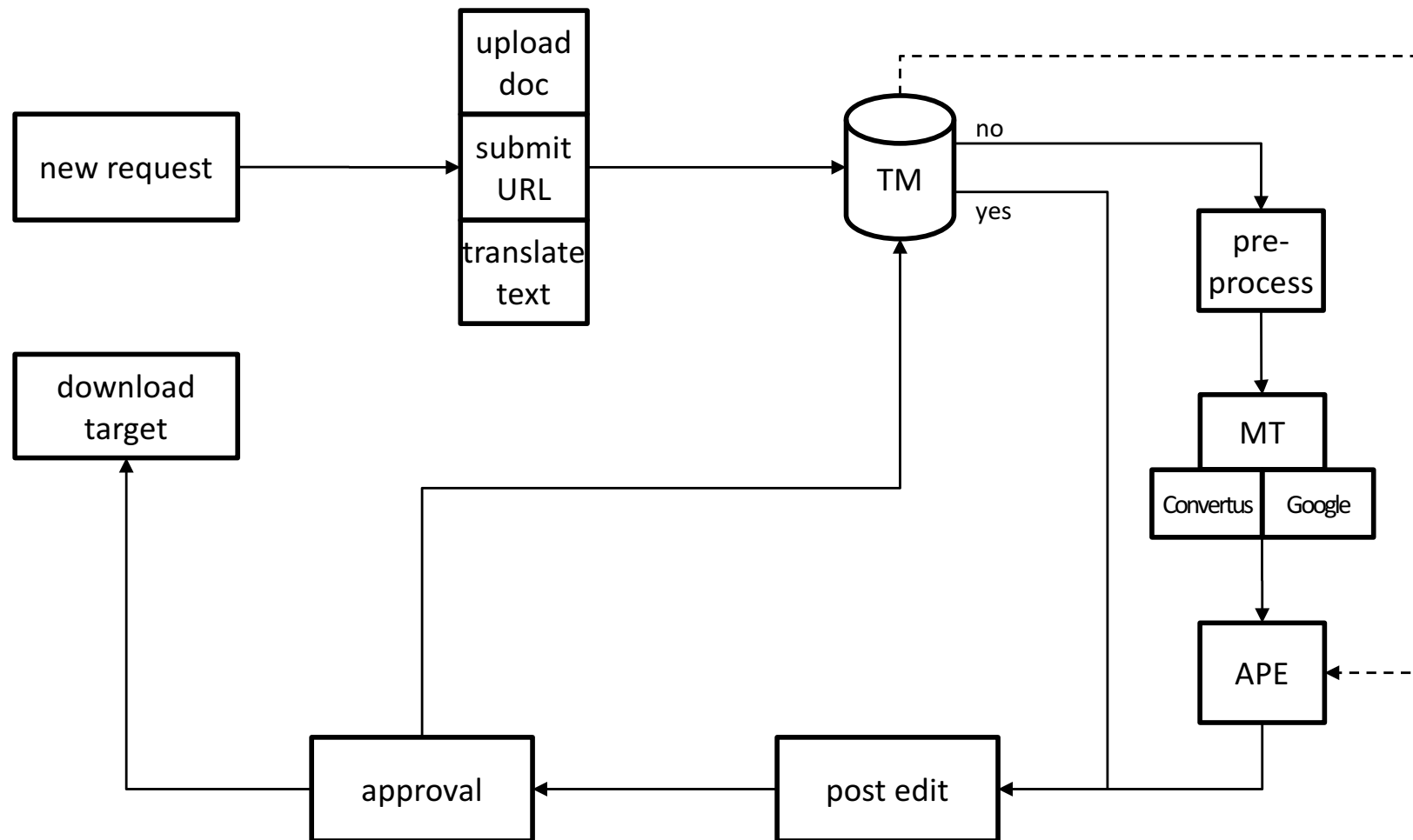


Convertus AB

Bakgrund

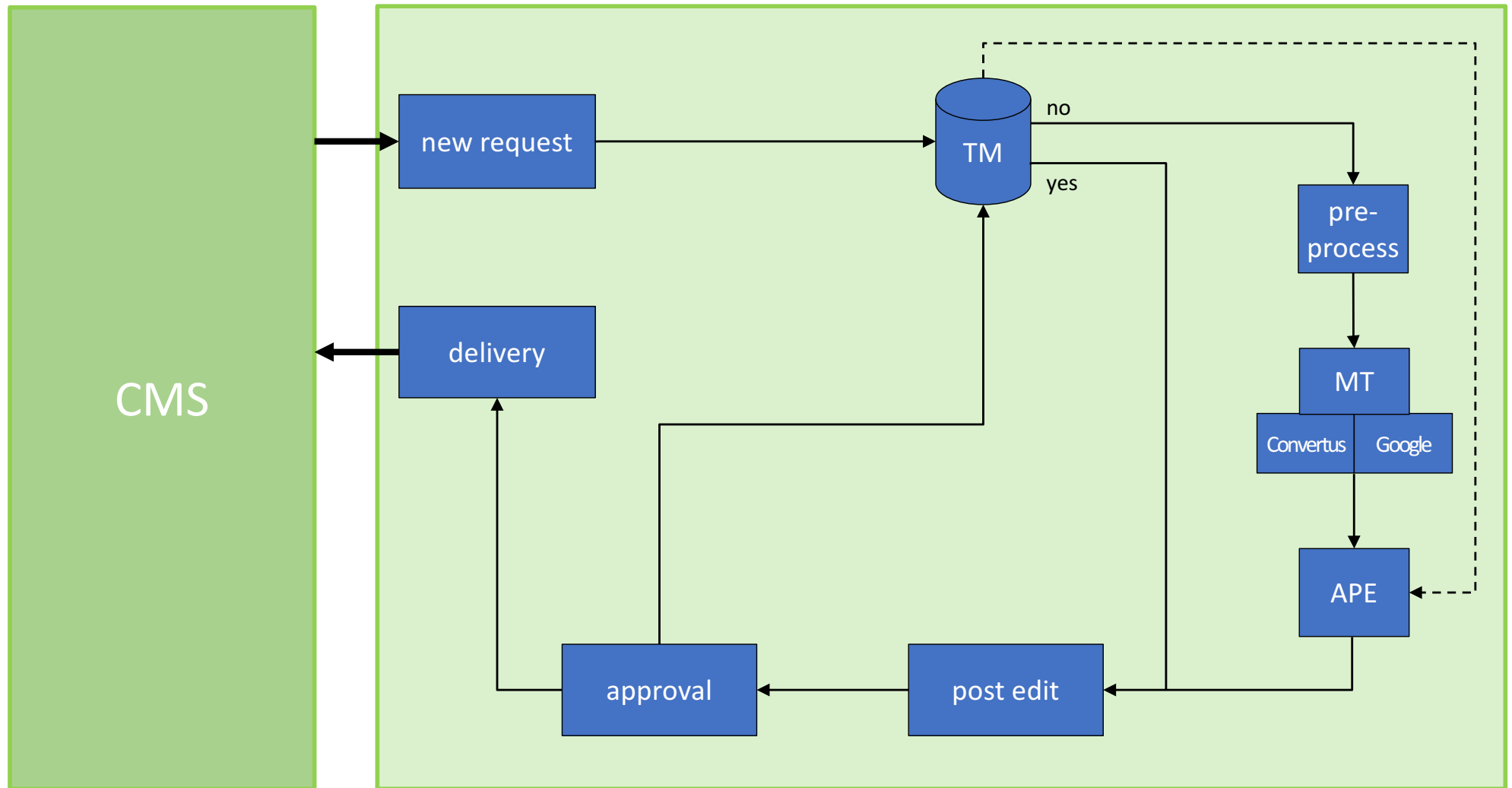
- Convertus roll i DigInclude är att utveckla och tillhandahålla översättningstjänster för översättning av användarorienterad myndighetsinformation
- Översättning från svenska till immigrantspråk
 - i första hand engelska, arabiska, persiska och finska
- Maskinöversättningsplattform

Convertus Bologna Translation Service



Convertus Bologna Translation Service

Integrerad i CMS



Översättningsminne	svenska-persiska
Totalt	9 738
Publicerade ÖS	3 265 (<i>LK: 1 820, 1177: 1 445</i>)
Postredigerade ÖS	6 473 (<i>LK: 2 686, 1177: 3 787</i>)

Översättningsminne	svenska-arabiska
Totalt	10 749
Publicerade ÖS	2 686 (<i>LK: 858, 1177: 1 828</i>)
Postredigerade ÖS	8 063 (<i>LK: 2 776, 1177: 5 287</i>)

MT med och utan TM

Redigeringsprotokoll : sv-ar

Ms	Titel	Antal segment	Poäng [%]	MT- motor	Poäng [%]	Minne	Poäng [%]
ar	Boende för äldre	31	69	16	62	15	81
ar	Boendestöd Team Socialpsykiatri	38	71	36	71	2	100
ar	Familjecentraler	42	84	25	78	17	100
ar	grundskola	32	72	19	66	13	100
ar	gymnasieskola	67	77	44	73	23	97
ar	Hemtjänst och trygghetslarm	40	81	26	74	14	100
ar	Kontakta kommunen om omsorg och hjälp	56	81	27	76	29	90
ar	Rehabilitering till arbete	40	60	39	60	1	100
		346	74	232	70	114	96

Poängberäkning:

$$\text{Poäng} = \frac{\text{Antal orörda ord}}{\text{Totalt ordantal}} \%$$

MT med och utan TM

Redigeringsprotokoll: sv-fa

Ms	Titel	Antal segment	Poäng [%]	MT- motor	Poäng [%]	Minne	Poäng [%]
fa	Boende för äldre	31	83	16	73	15	96
fa	Boendestöd Team Socialpsykiatri	38	66	36	65	2	96
fa	Familjecentraler	42	75	25	72	17	83
fa	grundskola	32	71	25	70	7	83
fa	gymnasieskola	67	74	47	71	20	90
fa	Hemtjänst och trygghetslarm	40	81	26	75	14	96
fa	Kontakta kommunen om omsorg och hjälp	56	73	26	63	30	90
fa	Rehabilitering till arbete	40	59	39	58	1	93
		346	73	240	68	106	91

Poängberäkning:

$$\text{Poäng} = \frac{\text{Antal orörda ord}}{\text{Totalt ordantal}} \%$$



MT med och utan TM - BLEU

Linköpings kommun

svenska–arabiska	BLEU
Ren maskinöversättning	38,0 %
Med minne (23 träffar)	43,1 %

svenska–persiska	BLEU
Ren maskinöversättning	38,9 %
Med minne (20 träffar)	45,6 %

Pre-editering

Kan omformulering av den svenska källtexten bidra till ökad översättningskvalitet?

Kan existerande förenklingslösningar, som till exempel StilLett, användas för datamaskinell pre-editering?

StilLett

StilLett (Rennes&Jönsson 2015) använder sig av 7 olika förenklingstyper:

- **Proximation:** ("man"->"du")
- **QuoteInvert:** ("Hej!" sa Kalle. -> Kalle sa: "Hej!")
- **StraightOrder:** (omvänd ordföljd -> rak ordföljd)
- **Pass2act :** (passiv->aktiv)
- **Decker:** (Decker 2003)
- **Split:** (Uppdelning av meningar vid konjunktion och subjunktionen "som"; duplicering av subjekt)
- **Syntax:** (Uppdelning av meningar vid konjunktion)

Rennes, Jönsson (2015): A Tool for Automatic Simplification of Swedish Texts

Decker

Decker (2003) är en förenklingstyp, som innebär att vissa element tas bort, t.ex. artiklar, som-satser och vissa prepositionsfraser.

Dessa typer påverkar sålunda innehållet och lämpar sig inte för pre-editering inför maskinöversättning.

Decker (2003): Towards automatic grammatical simplification of Swedish text.



Utvärdering av StilLett

Utvecklingskorpusen har körts genom StilLett och översatts till arabiska och persiska med Google (neuralt baserad). Vidare har BLEU-värden tagits fram för maskinöversättningen.

De olika typerna har körts var och en för sig och samlat.

För förenklingstypen omvänd ordföljd steg BLEU-värdet något för persiskan från 9,46 till 9,52. För arabiskan sjönk det istället från 10,75 till 10,36.

För övriga typer sjönk BLEU-värdet för båda språken. Det lägsta värdet (10,17 respektive 9,04) gav *split* (arabiska) och *Decker* (persiska). Med alla regler påkopplade sjönk BLEU-värdet till 9,36 för arabiska och 8,49 för persiska.

Se vidare, Stadell, Lina, 2017, Utvärdering av StilLett. Convertus. Intern projektrapport.

Tabell

Förenklingsmetod	Arabiska	Persiska
-	10,75	9,46
Proximation	10,63	9,26
quoteInvert	10,58	9,30
straightOrder	10,36	9,52
syntax	10,52	9,45
pass2act	10,27	9,37
decker	10,45	9,04
split	10,17	9,38
Alla metoder i kombination	9,36	8,49



Nya regler

Strängbaserade omskrivningsregler på otaggad text

Strängbaserade omskrivningsregler på taggad text

Parafrasering med Multra (Sågvall Hein 1997),
Convertus transferbaserade maskinöversättningssystem

Sågvall Hein, A., 1997, [Language Control and Machine Translation](#). I: Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation. July 23-25, 1997. St. John's College, Santa Fe, New Mexico.



Otaggade strängbaserad regler

- *hen -> han/hon*
- *månadens/periodens vara -> månadens/periodens produkt*
- *är med -> är närvarande*



Taggade, strängbaserade regler

Verb i presens i meningsinledning + *du* -> *Om du* + verb

Det är + nominalfras + *som* + verb -> nominalfras + verb

tillföra + substantiv -> *förse* + substantiv + *med*

preposition + substantiv + *finns det ...* -> *Det finns ...*
preposition + substantiv

de + substantiv + *som* -> substantiv + *som*

man (pronomen) -> *du*

Parafrasering med Multra

Initialt adverbial, omvänd ordföljd -> Rak ordföljd, finalt adverbial

Från sex månaders ålder kan du ge barnet läkemedel som innehåller ibuprofen.

Du kan ge barnet läkemedel som innehåller ibuprofen från sex månaders ålder.

Konditionalsats utan bisatsinledare -> Lägg till bisatsinledare

Har du många olika kontakter inom hälso- och sjukvården kan du be om att få en fast vårdkontakt.

Om du har många olika kontakter inom hälso- och sjukvården kan du be om att få en fast vårdkontakt.

Utelämnat hjälpverb i supinumformer -> Lägg till "har"

Om det inte känns bra kan du vända dig till verksamhetschefen för enheten där du fått vård.

Om det inte känns bra kan du vända dig till verksamhetschefen för enheten där du har fått vård.



Utvärdering mot testkorpus

1000 översättningssegment

	Arabiska	Persiska
Baseline	10,84	9,08
Strängbaserade regler	10,87	9,06
Strängbaserade regler med taggning	10,91	8,91
Multra	10,80	9,00
Strängbaserade regler + regler med taggning	10,94	8,90
Strängbaserade regler + taggning + Multra	10,85	8,87



Våra korpusar

- *Grundkorpus*
1767 publicerade översättningssegment på svenska, engelska, arabiska och persiska
- *Utvecklingskorpus* 500 översättningssegment samplade från grundkorpusen
- *Testkorpus* 1000 översättningssegment extraherade från grundkorpusen

Meningslängd

Strängbaserade omskrivningsregler på otaggad text, ingen effekt

BLEU-värden för alla olika meningslängder i en text – en splittrad bild

BLEU baserat på meningslängd

Antal ord	Antal meningar	BLEU	Antal ord	Antal meningar	BLEU
1	22	21,5	19	18	13,6
2	8	10,0	20	19	11,7
3	13	35,0	21	18	10,1
4	13	12,4	22	14	6,2
5	15	6,5	23	9	10,0
6	18	2,2	24	9	7,2
7	21	8,7	25	6	12,9
8	23	9,7	26	10	9,0
9	22	11,0	27	3	4,4
10	16	8,5	28	7	4,7
11	30	10,1	29	1	18,9
12	26	9,6	30	1	3,7
13	24	8,4	31	4	6,0
14	28	8,1	32	2	12,4
15	15	10,2	33	1	16,4
16	19	8,9	34	6	6,9
17	32	8,0	35	4	9,8
18	21	14,5	36	2	7,3



Uppföljning

Manuell utvärdering av testresultaten

Analys av vilka enskilda regler som givit positiva
översättningsresultat

Fortsatt regelutveckling och testning

Frågor?

