Eksempelbaseret maskinoversættelse

Rapport fra en workshop om EBMT

Hanne Fersøe

15. november 2001

Center for Sprogteknologi

hanne@cst.dk

EBMT på MT Summit 2001

Workshop på 8 timer med 8 foredrag

- **Ikke undervisning, dvs. ikke samlet overblik**
- **Diskussionsbidrag fra forskere til en igangværende debat**
- **Dvs. denne workshop var ikke særlig brugerorienteret**
- **Ingen produktionssystemer er i dag baseret på en EBMT-metode**

EBMT her i dag

Overblik over mit foredrag

1. **Baggrund for og overblik over EBMT**
2. **Gennemgang af væsentlige problemstillinger og konklusioner i workshoppens foredrag**
3. **Konklusion**

CST
CENTER FOR SPROGTEKNOLOGI

Baggrund for EBMT 1

Der begynder at findes og være adgang til store multilinguale korpora af oversatte tekster

Siden begyndelsen af 90'erne er man begyndt at bruge dem til MT (maskinoversættelse)

Man taler i dag om tre typer MT:

1. Regelbaseret MT (traditionel metode)

   To-sprogede ordbøger og grammatikker med syntaktisk og evt. semantisk viden

   Transferregler

Baggrund for EBMT 2

2. Statistikbaseret MT (bruger korpora)

kræver træning af systemet på meget store korpora af høj kvalitet

har endnu ikke vist høj kvalitet i komplekse overs.

3. Eksempelbaseret MT (bruger korpora)

udnytter og integrerer resurser (lingvistiske og statistiske) og teknikker (symbolske og numeriske)

TM anses for at være an variant af EBMT

Kombination af metoder: hybride systemer

CST
CENTER FOR SPROGTEKNOLOGI

Klassisk MT-arkitektur

- Regelbaserede og eksempelbaserede metoder har fælles grundlæggende arkitektur, og oversættelsen skabes af systemet
- Oversættelseshukommelse (TM) adskiller sig ved ikke at have en fase hvor der skabes en oversættelse af systemet, dette overlades til mennesket

Klassisk MT-arkitektur

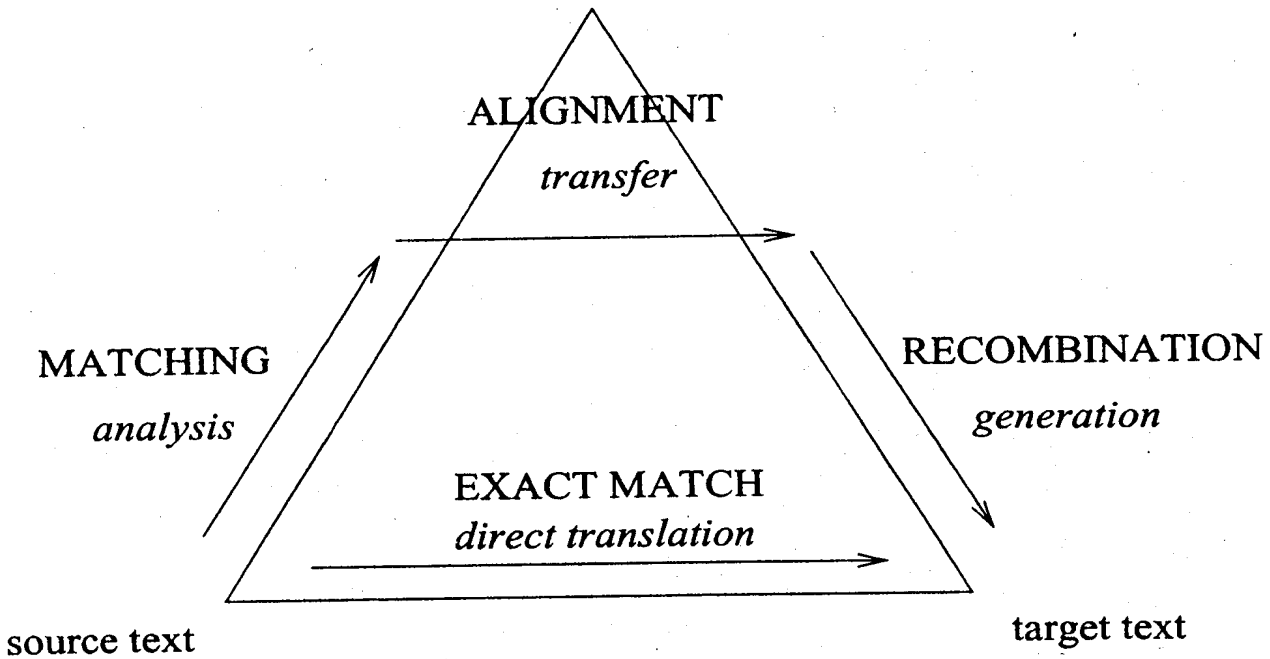


Figure 1. The "Vauquois pyramid" adapted for EBMT. The traditional labels are shown in *italics*; those for EBMT are in CAPITALS.

CST
CENTER FOR SPROGTEKNOLOGI

Overblik over EBMT 1

Fremskaffelse af parallelle korpora

- Det er blevet forholdsvis let
- Problemet er 'alignment' (sidestilling)

'Kornstørrelsen' på eksemplerne

- Den intuitive størrelse er sætningen
- Mennesker arbejder antagelig med mindre enheder
- Til 'matching' og 'recombination' behøves mindre stumper end sætninger (eks. 1)

Antallet af eksempler og kvalitet

- Jo flere eksempler jo bedre kvalitet
- Flere eksempler vil kun forbedre til en vis grænse
- For mange eksempler kan forværre

CST
CENTER FOR SPROGTEKNOLOGI

Overblik over EBMT 2

Eksemplernes egnethed

- Et 'aligned' parallelt korpus indeholder overlappende eksempler: nogle vil bekræfte hinanden, andre vil være i modstrid
- Bekræftende eksempler er ikke altid en fordel - kan føre til overgenerering
- Indførelse af skelnen mellem undtagelseseksempler og generaliserende eksempler

Overblik over EBMT 3

Lagring af eksemplerne

- Som annoterede træstrukturer (eks. 2)
- Som generaliserede eksempler (mønstre eller opskrifter/modeller) (eks. 3)
- Som forudberegnede statistiske parametre til hhv. oversættelsesmodel og sprogmodel

Matching

- Tegnbaseret, fx traditionel 'pattern-matching'
- Ordbaseret, fx m. tesaurus eller lignende
- Ordbaseret, fx ud fra ordklasseannoteringer
- Strukturbaseret, fx træer
- Delvis match, dvs. opsplitning på fragmenter

Overblik over EBMT 4

Alignment og Recombination

De vanskeligste trin i EBMT:

1) at afgøre hvilken del af en given oversættelse svarer til de fundne dele af kildeteksten

2) at kombinere de givne oversættelsesdele på en passende måde til et velformet målsprog

Afgrænsningsproblemer (boundary frictions) opstår især ved sprog med mange syntaktiske indikatorer, fx tysk (eks. 4)

Nogle eksempler fra databasen er mere brugbare end andre, fx afhængig af konkret kontekst

Brug af forudberegnede modeller: oversættelsesmodel og sprogmodel i statistisk MT

Overblik over EBMT 5

Systemtekniske problemer

- Store omkostninger i form af algoritmer til
  - at skabe
  - at lagre
  - at matche
  - at fremdrage
  - ud fra dette synspunkt er det dyrt 'blot' at tilføje flere eksempler
- Hastighed - løses
  - ved parallelprocessering eller
  - ved at skabe generaliseringer - bl.a dette har ført til udviklingen af hybride systemer

Foredragene på MT-Summit

- ✓ Definitionskriterier for EBMT - 1 stk.
- ✓ EBMT i forhold til CBR - 1 stk.
- ✓ Fraseleksikon (PL) i hybrid model - 1 stk.
- Forskellige algoritmer og metoder til at tilføre sproglig information til eller udlede lingvistiske regler fra eksempelbasen - 4 stk.
- ✓ Systemgennemgang og -demo - 1 stk.

CST
CENTER FOR SPROGTEKNOLOGI

Davide Turcato and Fred Popowich

What is EBMT?

Diskussion af definitionskriterier for EBMT

- **Implicit vs. Eksplicit viden**
  - **Uddragning på forhånd eller uddragning under kørsel?**
  - **De siger: Metoden er ligegyldig; det vigtige er hvilken slags viden der anvendes**
- **Én vs. flere kilder til viden**
  - **Tradition: Eksempeldatabasen er hovedkilden**
  - **De siger: Antallet af kilder er ligegyldigt; det vigtige er hvilken slags viden der anvendes**

CST
CENTER FOR SPROGTEKNOLOGI

Davide Turcato and Fred Popowich

What is EBMT? - fortsat

Deres konklusion

- **Definitionskriterier**
  - **Korpus er en kilde til viden**
  - **Korpus er en måde at lagre viden i kompakt form**
  - **Korpus er et lager af viden som systemet skal bruge, og som kun kan lagres på denne form**
- **Desuden er det karakteristisk for EBMT**
  - **Korpus er kun én af flere typer videnresurser**
  - **De fleste typer komponenter har en modpart i konventionel MT**
  - **Translation by analogy er den eneste virkelig eksempelbaserede metode**

CST
CENTER FOR SPROGTEKNOLOGI

Harold Somers

EBMT seen as Case-based Reasoning

- **CBR går ud på maskinelt at ræsonnere løsningen frem på et problem ved at bruge eksisterende cases med tilhøren-de løsninger**
- **Han diskuterer**
  - **Om metoder og modeller brugt i CBR (en disciplin inden for AI (kunstig intelligens)**
  - **kan overføres til MT problemstillingen**
  - **og om denne nye anvendelse af CBR kan anvise bedre veje til EBMT**
- **Han konkluderer**

  At EBMT ikke kan lære noget af CBR-metoden

CST
CENTER FOR SPROGTEKNOLOGI

Reinhardt Schäler

Beyond Translation Memories

- **Udgangspunkt: Oversættelse som en profession og som en forretning**
- **Forholdet mellem TM og EBMT**
- **Begrænsning for TM**
  - **Fokus på segmenter er gammeldags**
  - **Sandsynligheden for at finde en eksakt match på fraseniveau er større end på det nuværende TM segmentniveau**
- **Der ligger i dag meget store resurser i eksisterende TMer bl.a. takket være de alignmentværktøjer der sælges med TM systemer**

CST
CENTER FOR SPROGTEKNOLOGI

Reinhardt Schäler

Beyond Translation Memories

- **Forslag**
  - **Man skal bruge fraseleksikoner, PL (Phrasal Lexicon) på både kilde- og målsprog**
  - **Man skal udvikle automatiske eller semi-automatiske metoder til produktion af to-sproglige PLer med lingvistiske annotationer**
  - **Kendt hybrid metode: kombination af TM og MT foreslås udvidet med et PL til håndtering af situationer hvor der kun findes en fuzzy match på segmentniveau**
- **Har deltaget i at udvikle en første demonstrator til sådan et system**

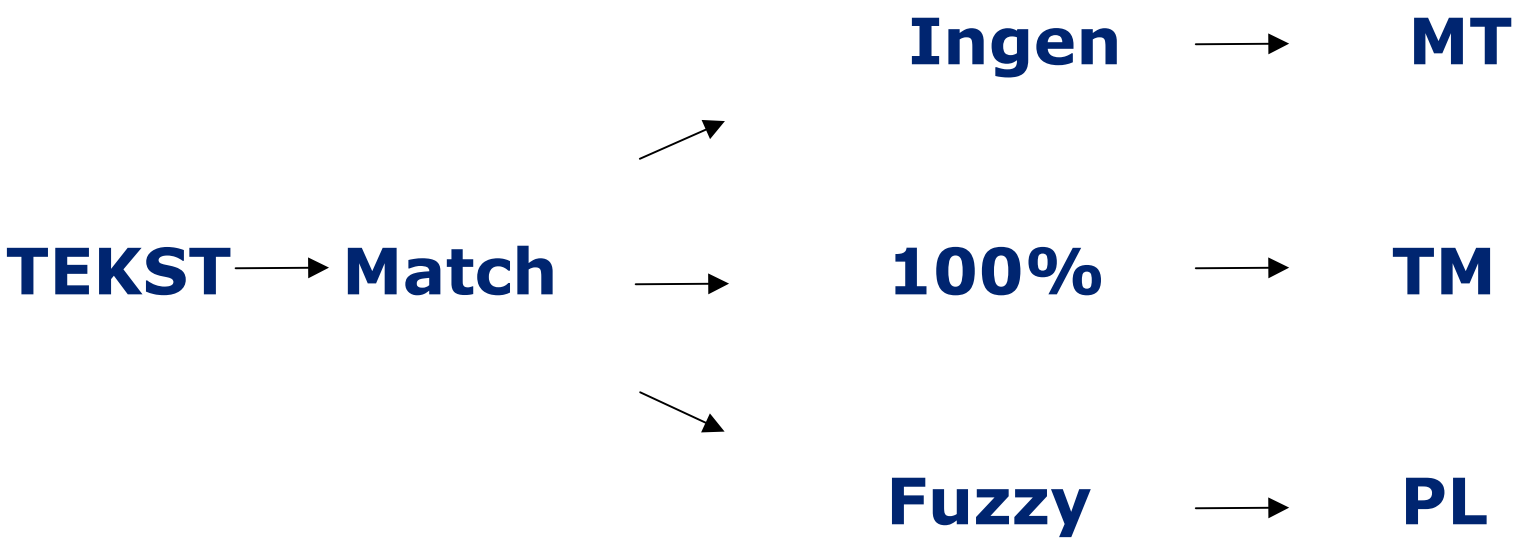
CST
CENTER FOR SPROGTEKNOLOGI

Eksempel

- **To TM indgange**
  - **1**
  - **[eng] The bullets move to the new paragraph**
  - **[ty] Die Blickfangpunkte rücken in den neuen Abschnitt**
  - **2**
  - **[eng] The title moves to the centre of the slide**
  - **[ty] Der Titel rückt in die Mitte des Dias**
- **Ny sætning der ikke kan oversættes automatisk ud fra disse TM indgange**
  - **The bullets move to the centre of the slide**

EUROMAP LANGUAGE TECHNOLOGIES

En udvidet hybrid model



Arul Menezes and Stephen D. Richardson

A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora

- **Introduktion**
- **Et MT-system kræver en omfattende mængde viden om oversættelse. Viden ligger i fx ordbøger, oversættelsesregler, sprogmodeller, eksempeldatabaser, statistiske regler osv.**
- **Man vil gerne finde en måde til at få automatisk adgang til denne viden ud fra bilinguale korpora**
- **Nogle systemer (statistiske) bygger oversættelsesmodeller med denne viden uden sproglig analyse**
- **Andre systemer, fx det her præsenterede, parser sætninger i korpora der er parallelalignede på sætningsniveau. Resultatet er strukturer på de to sprog. Strukturerne alignes herefter og man udleder såkaldte transfer-mappings - oversættelsesregler - som udgør systemets eksempeldatabase. (eks.5)**

CST
CENTER FOR SPROGTEKNOLOGI

Vanskelighederne i dette

- **Alignmentproceduren**
  - **Proceduren må være meget robust overfor parserfejl og overfor fejl den selv producerer**
- **Udledningen af transfer-mappings, dvs. ækvivalenter**
  - **Den må producere meget præcise regler**
  - **Reglerne må have tilstrækkelig kontekst-information til at gøre systemet i stand til at vælge en passende oversættelse.**

Arul Menezes and Stephen D. Richardson

A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora

- **Resten af artiklen handler om**
- **Algoritmen til alignment**
- **Algoritmen til transfer mappings**
- **Rapport om eksperimenter og resultater**

  1. Sammenligning af resultater opnået med best-first algoritmen i forhold til BabelFish og 3 andre algoritmer i deres eget system:

  Fem evaluatorer, blind test, sammenligning af MT-oversættelser, reference: en manuel oversættelse, ingen kildetekst. De skulle vælge hvilke oversættelse der var bedst (eks 6.)

CST
CENTER FOR SPROGTEKNOLOGI

2. Absolut kvalitet - sammenligning af resultater opnået med best-first og med BabelFish

Fem evaluatorer, blind test, sammenligning af MT-oversættelser, reference: en manuel oversættelse, ingen kildetekst. De skulle vurdere hver maskinoversættelse på en skala fra 1-4, hvor 4 er den ideelle oversættelse og 1 uacceptabel (eks. 7)

- Konklusion
- De er meget tilfredse med resultaterne
- De vil arbejde videre med teknikker til at inkludere mere kontekst for at blive bedre til at disambiguere mellem mappings som er i konflikt med hinanden (teknikker fra Machine Learning)

Rapport fra en workshop om EBMT

- **Konklusion**
- **Hele MT-samfundet virker meget indstillet på at kombinere metoder og teknikker på kryds og tværs for at opnå bedre oversættelsesresultater**
- **Arbejdet kræver store korpora så for at mindre udbredte sprog som dansk også skal kunne få glæde af disse metoder og teknikker må forskerne have adgang til korpora.**