Statistilise masintõlge väljundi parandamine

Mark Fišel (fishel@ut.ee)

01. märts 2007

Ettekanne sisu

- statistiline masintõlge (SMT);
- SMT seis Eesti keele puhul;
- olemasolevad parandamise meetodid;
- alternatiivne parandamise meetod

Statistiline masintõlge

$$e,\ f,\ p(e),\ p(e|f)$$

$$\begin{aligned} \hat{e} = \arg\max_e p(e|f) &= \arg\max_e \frac{p(e)p(f|e)}{p(f)} \\ &= \arg\max_e p(e)p(f|e) \end{aligned}$$

- $p(f|e)$ – tõlkimismudel
- $p(e)$ – keelemudel
- arg max – dekodeerija

Fraasipõhine TM

Fraasipõhises tõlkimismudelis modelleeritakse erinevad etapid sõltumatult:

- fraaside tõlkimine
- fraaside järjekord

Probleemid:

- enne nägematud sõnad
- valesti valitud sõnad/fraasid
- vale fraaside järjekord

SMT + Eesti keel

- Eesti-Inglise suund
- testitud 2 korpuse peal, mõlemad – seadustekstidega
- parim tulemus:
  - BLEU: 45.22%
  - käsitsi hinnang:

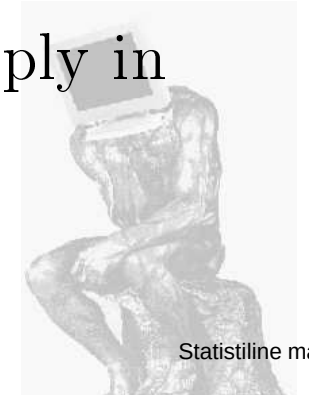
Hea	15%
Rahuldav	15%
Vale	70%

Eesti-Inglise SMT probleemid

- hõredad andmed (sparse data)
  - mitu korda sisendsõna jääb tõlkimatuks – see ei esinenud treenimiskorpuses
  - seda põhjustab väike korpuste suurus, eesti keele rikas morfoloogia, liitsõnad, jne
- vale järjekord
  - järjekorra komponent ei saa hakkama

Vale järjekord

—(euroopa majandusühenduse ja šveitsi konföderatsiooni)$_1$ (vaheliste kokkulepete)$_2$ (kohaldamisel)$_3$ (rakendatakse ühenduses)$_4$ (ühiskomitee)$_5$ (otsust nr 5 / 81)$_6$
—(for the purposes of application)$_3$ (of the agreements between)$_2$ (the european economic community and the swiss confederation)$_1$, (decision no 5 / 81)$_6$ (of the joint committee)$_5$ (shall apply in the community)$_4$
—(the european economic community and the swiss confederation)$_1$ (of agreements between)$_2$ (the application)$_3$ (of the joint committee)$_5$ (shall apply in the community)$_4$ (decision no 5 / 81)$_6$



Olemasolev parandamise meetod

- kasutades teadmisi kahe keele grammatikat saab enne treenimist ja tõlkimist muuta sisend-keele järjekorda, et see oleks rohkem väljund-keelega sarnane, või
- järjekorra muutmise mustreid saab leida käsitsi või masinõppimisega
  - käsitsi tegemine nõuab väga palju inim- ja ajaresursse
  - masinõppe nõuab süntaksanalüsaatorit mõlema keele jaoks

Alternatiivne parandamise meetod

- oletame et me saame tõlkida sõnade asemel sõnaliike, ja seda väga hea täpsusega
  - S V A $\Rightarrow$ NN VBZ JJ
- siis saab tõlkida enne lause sõnaliigid, siis lauset ennast, ja siis muuta lause fraaside järjekorda vastavalt sõnaliikidele

```
kommitee otsus nr 5 ->
     of the commitee decision nr 5
     IN DT  NN        NN        NN CD

S S S N -> NN NN CD IN DT NN

of the commitee decision nr 5 ->
     decision nr 5 of the commitee
```

Sõnaliikide tõlkimine

- peaks olema lihtsam, kuna sõnavara on piiratud (eesti – 16, inglise Penn – umb. 50), mingit morfoloogiat või liitsõnu pole
- praktikas aga tavaline SMT töötab sõnaliikidega isegi halvemini kui sõnadega
  - üks teatud probleemide hulgast: peaaegu igas lauses on nii tegu- kui nimisõna, kuidas teha vahet?

Järjekorra muutmine

- brute force: $O(n!)$, kus n on sõnade (fraaside) arv
- mis võiks olla effektiivsem algoritm?
- probleemi teeb lihtsamaks see et väljund on juba jagatud fraasideks
- aga:
  - sõnaliigid võivad korduda, eriti nimisõnafraasides (lepingu allkirjastamise kuupäev – S S S)
  - tõlgitud variant ei ole ideaalne, ei pruugi sisaldada kõike "oraakli" sõnaliike

Kokkuvõte

- SMT
- Eesti SMT
- vale järjekord
- parandamise meetodid

Aitäh!