

Masintõlge 1, 2

Heiki-Jaan Kaalep, TÜ

Momendi seis

- <http://translate.google.com/>
- <http://ats.cs.ut.ee/smt/translate/>
- Nii Google'i kui TÜ versioon kasutavad statistilist masintõlkemeetodit
- Tegijad: Mark Fišel, Harri Kirik, Katrin Tsepelina, Kadri Muischnek, Kaarel Veskis

Projekti senised tulemused

- Korpused + tarkvara + kohandamine = demo
- korpused
 - korpuste tegemine, hindamine, kombineerimine ja kvaliteedi parandamine
- tarkvara
 - Moses (keelest sõltumatu, vaba)
- kohandamine
 - eesti keele morfoloogia (mida maksab arvestada), katsed lingvistilise töötlusega

Olemasolevad korpused

- Õigustekstid
 - TÜ paralleelkorpus: 8 mln sõna inglise, 5 mln – eesti keeles
 - Ispra JRC-Acquis: 55 mln sõna inglise, 40 mln – eesti keeles
- Subtiitrid
 - Opus korpus
- KDE

Ressursside kombineerimine

- Sõltuvalt kasutatud meetodist eri korpustes 85-95% korrektsest paralleeliseeritud lauseid
 - Parim meetod - HunAlign
- Automaatne kombineerimine
 - arvestab väikeste erinevustega korpuste lausetes
 - arvestab korpuste erinevate paralleeliseerimise tasemetega
- Tulemuseks saadud korpus korrektsem ja suurem kui algkorpused
- Suurem korpus, paremad tulemused

Kohandamine

- Katsed erinevate korpuste ja erinevate sõnaanalüüsi viisidega on näidanud, et:
 1. Eestikeelsete sõnade morfoloogiline analüüs, milles sõnad tükeldatakse tüvedeks ja lõppudeks, aitab kaasa õigete ingliskeelsete fraaside leidmisele.
 2. See ei aita parandada tõlkeprobleeme, mille põhjuseks on eesti ja inglise keele erinev sõnajärg.

Harri Kiriku bakalaureusetöö:

<http://math.ut.ee/~harts/thesis.html>

Näited

- Sünonüümid
 - millal maksan eide vaeva ?
- Sõnade järjekord
 - ei karda eesti rind
 - pangale võlgu jäävate inimeste hulk teeb valitsusele peavalu
- Korpuses sageli esinenud väljendid (aga mitte sõna-sõnalt samad)
 - Eesti kavatseb oma otsused kiiresti vastu võtta

Lisaks: leksikonid paralleelkorpuste baasil

- Leksikoni tuletamine (K. Tsepelina, K. Veskis)
 - rakendatud KDE korpusele, et luua arvutialase terminoloogia kasutamist kajastav leksikon
 - plaan: võrrelda muude arvutileksikonidega keeleveebis
- M. Traadi projekt (2009-2010)

Edaspidine töö

- parandada kvaliteeti
 - keelespetsiifiline tarkvara, eeskätt süntaksi analüüs
 - demo tagasiside ja alternatiivide võrdlemine
 - kombineerida erineval moel treenitud programmiversioone ja valida väljund mitme variandi hulgast (hääletamine)

Aitäh!