T-(538|725)-MALV, Málvinnsla
Kynning

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]Tölvunarfræðideild, Háskólinn í Reykjavík

Ágúst 2007

Outline

Outline

Markmið og lýsing

Námsmarkmið

- Að nemendur þekki helstu aðferðir sem notaðar eru á sviði málvinnslu.
- Að nemendur kynnist helstu rannsóknum sem fengist er við á sviðinu.
- Að nemendur geti útfært kerfi sem vinnur með náttúrulegt tungumál.

Lýsing

Máltækni (tungutækni) er svið sem hefur það að markmiði að smíða kerfi sem gera fólki kleift að eiga samskipti við tölvur með því að nota náttúrlegt tungumál. Máltækni er þverfaglegt svið sem krefst þekkingar á sviðum eins og málvísindum, tölfræði, sálfræði, verkfræði og tölvunarfræði. Í þessu námskeiði er fjallað um grundvallaratriði í málvinnslu ... (sjá frekar á vefsíðu námskeiðs)

Markmið og lýsing

Námsmarkmið

- Að nemendur þekki helstu aðferðir sem notaðar eru á sviði málvinnslu.
- Að nemendur kynnist helstu rannsóknum sem fengist er við á sviðinu.
- Að nemendur geti útfært kerfi sem vinnur með náttúrulegt tungumál.

Lýsing

Máltækni (tungutækni) er svið sem hefur það að markmiði að smíða kerfi sem gera fólki kleift að eiga samskipti við tölvur með því að nota náttúrlegt tungumál. Máltækni er þverfaglegt svið sem krefst þekkingar á sviðum eins og málvísindum, tölfræði, sálfræði, verkfræði og tölvunarfræði. Í þessu námskeiði er fjallað um grundvallaratriði í málvinnslu ... (sjá frekar á vefsíðu námskeiðs)

Kennslubækur

Aðal kennslubók

An Introduction to Language Processing with Perl and Prolog

Ítarefni - aðgengilegt á bókasafni HR

- Foundations of Statistical Natural Language Processing.
- Speech and Language Processing.
- Handbook of Natural Language Processing.
- Learning Perl.
- Annað efni er sótt af vefnum eða ljósritað.

Kennslubækur

Aðal kennslubók

An Introduction to Language Processing with Perl and Prolog

Ítarefni - aðgengilegt á bókasafni HR

- Foundations of Statistical Natural Language Processing.
- Speech and Language Processing.
- Handbook of Natural Language Processing.
- Learning Perl.
- Annað efni er sótt af vefnum eða ljósritað.

Námsmat

Vægi einstakra þátta

- Forritunarverkefni: 40%
- Skriflegt lokapróf: 30%
- Þátttaka í námskeiði: 15%
- Skiladæmi: 15%

Annað

- Verkefni eru einstaklingsverkefni.
- Nemandi þarf að ná lokaprófi til að standast námskeiðið.
- Verkefnaáætlun má finna á enskri vefsíðu námskeiðsins, `http://cadia.ru.is/wiki/public:t-malv-07-3:main`.

Námsmat

Vægi einstakra þátta

- Forritunarverkefni: 40%
- Skriflegt lokapróf: 30%
- Þátttaka í námskeiði: 15%
- Skiladæmi: 15%

Annað

- Verkefni eru einstaklingsverkefni.
- Nemandi þarf að ná lokaprófi til að standast námskeiðið.
- Verkefnaáætlun má finna á enskri vefsíðu námskeiðsins, `http://cadia.ru.is/wiki/public:t-malv-07-3:main`.

Forritunarverkefni

Samanstendur af fimm hlutum

- 1. hluti: Tilreiðing texta
- 2. hluti: Mörkun texta
- 3. hluti: Þáttun texta
  - BS-nemendur fá þennan hluta "gefinn"
- 4. hluti: Orðræðulíkan
- 5. hluti: Heildarkerfi

Þátttaka í námskeiði og skiladæmi

Þátttaka

- Fræðigrein lesin og kynnt í fyrirlestri.
- Þátttaka í umræðuþráðum utan fyrirlestra.

Skiladæmi

- Þrjú skiladæmi verða lögð fyrir.

Þátttaka í námskeiði og skiladæmi

Þátttaka

- Fræðigrein lesin og kynnt í fyrirlestri.
- Þátttaka í umræðuþráðum utan fyrirlestra.

Skiladæmi

- Þrjú skiladæmi verða lögð fyrir.

Outline

Hvað er máltækni?

Skilgreining

- Rannsóknar- og þróunarsvið þar sem fengist er við að smíða kerfi sem geta unnið með og skilið náttúruleg tungumál og stuðlað að notkun þeirra í samskiptum manns og tölvu.
- Enska heitið er "(Human) Language Technology" (HLT eða LT).
- Þverfagleg fræðigrein — samspil málvísinda, sálfræði, tölfræði, tölvunarfræði, verkfræði.

Tvö undirsvið

- Textavinnsla (e. Text (Language) Processing)
- Talvinnsla (e. Speech Processing)

Hvað er máltækni?

Skilgreining

- Rannsóknar- og þróunarsvið þar sem fengist er við að smíða kerfi sem geta unnið með og skilið náttúruleg tungumál og stuðlað að notkun þeirra í samskiptum manns og tölvu.
- Enska heitið er "(Human) Language Technology" (HLT eða LT).
- Þverfagleg fræðigrein — samspil málvísinda, sálfræði, tölfræði, tölvunarfræði, verkfræði.

Tvö undirsvið

- Textavinnsla (e. Text (Language) Processing)
- Talvinnsla (e. Speech Processing)

Málvinnsla

Máltækni vs. málvinnsla

- "Language Technology" (LT) $\approx$ "Natural Language Processing" (NLP)
- Máltækni $\approx$ málvinnsla
- Málvinnsla leggur áherslu á:
  - Greiningu (e. analysis) formgerðar (og merkingar) máls.
  - Myndun (e. generation) máls út frá formgerð (merkingu).
- NLP $\approx$ "Computational Linguistics" (tölvufræðileg málvísindi)

Outline

Máltækniverkefni

Dæmi

- **Málfræðileiðréttingar** (e. grammar checkers)
  - `http://en.wikipedia.org/wiki/Grammar_checker`
- **Upplýsingaheimt** (e. information retrieval) og **upplýsingaútdráttur** (e. information extraction)
  - `http://en.wikipedia.org/wiki/Information_extraction`
- **Fyrirspurnarkerfi** (e. Question-Answering Systems)
  - `http://en.wikipedia.org/wiki/Question_answering`
- **Vélrænar þýðingar** (e. Machine Translation)
  - `http://en.wikipedia.org/wiki/Machine_Translation`

Máltækniverkefni

Fleiri dæmi

- **Talkennsl/Talgreining** (e. Speech recognition)
  - `http://en.wikipedia.org/wiki/Speech_recognition`
- **Talgerving** (e. Speech synthesis; text-to-speech)
  - `http://en.wikipedia.org/wiki/Speech_synthesis`
- **Samræðukerfi** (e. Dialogue Systems)
  - `http://nlp.shef.ac.uk/research/areas/dialogue.html`

Talandi vélmenni

HAL

- Kvikmyndin 2001: Space Odyssey. Leikstjóri: Stanley Kubric; gerð árið 1968.
- Tölva sem talar og skilur ensku.
- Myndin spáði 33 ár fram í tímann.
- Hversu nálægt er spáin raunveruleikanum?
- Hvað þarf til að búa til "veru" sem býr yfir málmyndun og málskilningi HALs?

Outline

Fræðigreinar málvísinda

- Hljóðfræði og hljóðkerfisfræði (e. Phonetics and Phonology)
- Orðhlutafræði (e. Morphology)
- Setningafræði (e. Syntax)
- Merkingarfræði (e. Semantics)
- Orðræða og samræða (e. Discourse and Dialogue)

Þessar fræðigreinar mynda jafnframt mismunandi stig málvinnslu.

Outline

Margræðni (e. Ambiguity)

- Margræðni á sér stað þegar tiltekið inntak á sér mismunandi málfræðilegar formgerðir (e. linguistic structures), þ.e. hefur mismunandi merkingar í för með sér.
- Í flestum tilvikum leysa manneskjur margræðnina á ómeðvitaðan hátt.
- Margræðni "þjakar" hins vegar sérhvert stig í málvinnslu.
- Margræðni er eytt með einræðingu (e. disambiguation).

Margræðni í talgreiningu

Dæmi

- Inntak: The boys eat the sandwiches.
- Mögulegt úttak:
  - The boy seat the sandwiches.
  - The boy seat this and which is.
  - The boys eat this and which is.
  - The boys eat the sand which is.
  - etc.

Margræðni í orðflokksmörkun (e. part-of-speech tagging)

Dæmi

- Inntak: Hann á við.
- Mörk einstakra orða:
  - Hann=fpken_fpkeo
  - á=aþ_ao_sfg1en_sfg3en_aa_nven_nveo_nveþ
  - við=ao_fp1fn_aþ_aa_nkeo

Dæmi um merkingu einstakra stafa:

```
n=nefnifall, o=þolfall, þ=þágufall, e=eignarfall
n=nafnorð, f=fornafn, p=persónufornafn, a=atviskorð, s=sögn
k=karlkyn, v=kvenkyn
e=eintala, f=fleirtala
f=framsöguháttur, g=germynd
```

Margræðni í setningagreiningu/merkingargreiningu

Dæmi

- Inntak: I saw the boy with the telescope.
- Merking:
  - I used a telescope to see the boy.
  - I saw the boy who had a telescope.

Val og útfærsla á líkani (e. model)

Þegar náttúrulegt tungumál er greint þá:

- Þarf að búa til eða velja einhvers konar (formlegt) líkan.
  - Að búa til gott líkan er erfitt.
  - Mál er tengt mannlegri hugsun og skilningi.
- Og útfæra líkanið í forriti.

Outline

Tungutækni - Skýrsla starfshóps

Menntamálaráðuneytið, 1999

- `http://www.tungutaekni.is/news/Skyrsla.pdf`
- Grundvallarspurning: “Hví skyldi fámenn þjóð hafa fyrir því að leggja í það ærinn kostnað að gera tungumál sitt hæft til notkunar í alþjóðlegu upplýsingaþjóðfélagi?”
- Átak var lagt til á fjórum sviðum:
  - Byggð verði upp sameiginleg gagnasöfn, málsöfn, sem geti nýst fyrirtækjum sem hráefni í afurðir.
  - Fé verði veitt til að styrkja hagnýtar rannsóknir á sviðið tungutækni.
  - Fyrirtæki verði styrkt til þess að þróa afurðir tungutækni.
  - Menntun á sviði tungutækni og málvísinda verði efld.

Spjall vikunnar

Tungutækniverkefnið

- Í kjölfar skýrslunnar varð til Tungutækniverkefni menntamálaráðuneytisins sem stóð til 2004.
- Hver varð afrakstur þess og hvað hefur gerst síðan?
- Það er ykkar að finna út og ræða um í spjalli vikunnar.

Spjallþræðir

- `http://malv2007.proboards50.com/`

Outline

Vefir og demó

- Tungutæknisetur: `http://www.tungutaekni.is`
- Gervigreindarsetur HR: `http://ailab.ru.is/`
- IceNLP: `http://nlp.ru.is`
- Tungumálakennsla
- Hreyfimyndagerð