

Merkingarbrunnur fyrir íslenska máltækni

Matthew Whelpton
University of Iceland
whelpton@hi.is

Anna Björk Nikulásdóttir
University of Iceland
abn@hi.is



[Semantics in IT and NLP]

- Semantic information is an essential resource for Natural Language Processing applications, especially for:
 - information extraction
 - question answering
 - machine translation
 - integrated spoken language systems



[First steps for Icelandic]

- Three examples of semantic resources for English
 - WordNet
 - FrameNet
 - PropBank
- Overview of the first project aiming to develop a comparable resource for Icelandic
 - WordNet-like semantic network
 - RANNÍS



[WordNet]

- WordNet is
 - an electronic database
 - of (English) words
 - and the semantic relations between them



[Hyponymy]

- “fox” is_a_kind_of “animal”
 - every fox is an animal
 - NOT every animal is a fox
 - “fox” is a hyponym of “animal”
 - hypo ‘under’
 - nym ‘name’
 - “animal” is a hypernym of “fox”
 - hyper ‘over’
 - nym ‘name’



[Two simple uses]

- Helps with semantically sensitive searches
 - find animals that eat eggs
 - WordNet tells us that “fox” as an “animal” meets the first criterion of this search
- Helps with reference tracking
 - The farmer saw a fox. The animal was hurt.
 - => the fox was hurt



[FrameNet]

- WordNet focuses on individual words and semantic relations between them.
- FrameNet focuses on generalised situations (frames) and the entities and activities which characterise those situations (frame elements)



[Apply_heat Frame]

- Frame
 - Apply_heat
- Frame Elements
 - Cook
 - Food
 - Heating_Instrument
- Example
 - [*Cook* Matilde] [*Apply_heat* **fried**] [*Food* the catfish]
[*Heating_instrument* in a heavy iron skillet].



[Form and Meaning]

- WordNet
 - Individual words
 - Semantic relations between them
- FrameNet
 - General situations (word knowledge)
 - How words characterise those situations
- PropBank
 - Relation between
 - Syntactic Structure
 - Predicate Argument Structure
 - U.Penn. TreeBank + predicate argument information



Syntactic and Argument Structure

■ What do you like?

Treebank annotation:

(SBARQ (WHNP-1 what)
(SQ do
(NP-SBJ you)
(VP like
(NP *T*-1))))))

Propbank annotation:

Rel: like
Arg0: you
Arg1: [*T*] -> What

Roleset like.01

Arg0: experiencer of affection or preference
Arg1: object of affection or preference



[The Next Step (RANNÍS)]

- All of the resources mentioned were manually created
 - extremely labour-intensive
 - unrealistic for Icelandic
- Search for automated methods
- Anna Nikulásdóttir has already begun such work on the creation of a WordNet-like semantic database for Icelandic and this will be the focus of the new RANNÍS project.



[Greining merkingarvensla úr ÍO]

- Níu mismunandi merkingarvensl greind sjálfvirkt úr *Íslenskri orðabók* (ÍO)
- Notast við orðflokka- og blönduð orða- og orðflokka-myndir:

`l_n` → `yfirheiti(fletta, n)`
`eiginleiki(fletta, l)`

`það að_s` → `tengtSo(fletta, s)`



[Greining merkingarvensla úr ÍO]

fuglsblundur örstuttur (lo.) svefn (no.)
yfirheiti(fuglsblundur, svefn)
eiginleiki(fuglsblundur, örstuttur)

mæling það að mæla (so.)
tengtSo(mæling, mæla)



[Greining merkingarvensla úr ÍO]



[Greining merkingarvensla úr ÍO]

- Niðurstöður:
 - 77.348 skýringar eða 96,45% allra nafnorðaskýringa í ÍO greindar
 - Prófunarsett: 94,77% skýringa án rangrar greiningar



[Merkingarnám úr textum]

- Merkingarnám úr textum þarfnast gífurlegs magns markaðra texta
- Mörkuð íslensk málheild mun hér koma að miklum notum
- Stefnt er að því að finna mynstur í íslensku sambærileg svokölluðum Hearst-mynstrum í ensku



[Hearst-mynstur]

- NP_0 such as $NP_1, NP_2, \dots, NP_{n-1}$ (and/or) NP_n
... red algae, such as Gelidium, ...
- such NP_0 as $NP_1, NP_2, \dots, NP_{n-1}$ (and/or) NP_n
... works by such authors as Herrick, Goldsmith, and Shakespeare
- NP_1, NP_2, \dots, NP_n (and/or) other NP_0
Bruises, wounds, broken bones or other injuries ...
... temples, treasuries, and other important civic buildings



[Möguleg mynstur í íslensku?]

- NP_0 , utan $NP_1, NP_2, \dots, NP_{n-1}$ og NP_n
... heimilisstörf, utan eldamennsku ...
- NP_1, NP_2, \dots, NP_n eða (aðra|annan|annað|önnur) NP_0
... Diesel-gallabuxur eða aðra merkjavöru
- NP_0 eins og $NP_1, NP_2, \dots, NP_{n-1}$ (og|eða) NP_n
... grunngildum eins og góðum samskiptum ...



[Merkingarbrunnur?]

- Stefnt er að því að tengja niðurstöður innbyrðis
- Þróað verður tól með grafísku viðmóti til þess að leiðrétta og bæta við niðurstöður sjálfvirku greiningarinnar
- Möguleikar á tengingu við WordNet verða kannaðir

