IceParser: An Incremental Finite-State Parser for Icelandic

Daníel B. Sigurgeirsson

IceParser

- First parser for Icelandic
  - http://nlp.cs.ru.is/IceNLPWeb/icenlp.html
- Input: POS-tagged text
- Output: parsed text with phrases and syntatic structures marked
- Two-phase processing:
  - Phrase-structure module
  - Syntatic structure module
- Each phase is comprised of a series of finite-state transducers
  - Transducer is an automata that accepts, translates or generates a pair of strings

Shallow parsing

- Shallow parsing vs. Deep parsing:
  - Deep parsing builds a full parse tree for a given sentence, while shallow parsing only parses individual "chunks" of the sentence
- Benefits from using shallow parsing:
  - less complexity, more speed
  - more robust, parser is less sensitive to grammatical errors in the text, and/or low quality in the input (missing words, mistakes, noise)
  - works well when the language has a free word order (like Icelandic)
  - shallow parsing is sufficient for many applications
    - information extraction, text summarisation, grammar checking etc.

Reduction vs. Construction

- Reductionist method:
  - reduce all possible readings of a sentence (represented by finite-state automata) to one correct reading by a set of elimination rules.
- Constructive method:
  - based on a lexical description of a collection of syntactic patterns
- IceParser uses the constructive method
  - a sequence of transducers are chained together - forming a "pipeline"

Phase 1: Phrases

- The following phrases should be marked according to the EAGLES (Expert Advisory Group for Language Engineering Standards) standard:
  - AdvP (Adverb)
  - AP (Adjective)
  - NP (Noun)
  - PP (Preposition)
  - VP (Verb) - which are subclassified (VPx)
- Additionally, the following phrase categories are marked:
  - CP (Coordinating conjunction)
  - SCP (Subordinating conjunction)
  - InjP (Interjection)
  - APs (sequence of adjective phrases)
  - NPs (sequence of noun phrases)
  - MWE (Multi-word expressions)

Bottom-up method

- Phrases are marked using the bottom-up method:
  - AdvP are marked before AP,
  - AP are marked before NP
  - etc.
- Example:
  - mjög góður (very good)
  - [AdvP mjög AdvP] góður
  - [AP [AdvP mjög AdvP] góður AP]

Phase 2: Syntatic structures

- Curly braces denote a syntatic function
- The following tags are used:
  - *QUAL - (genitive qualifier)
  - *SUBJ - (subject)
  - *OBJ - (object)
  - *OBJAP - (object of an AP)
  - *OBJNOM - (nominative object)
  - *IOBJ - (indirect object)
  - *COMP - (complement)
  - *TIMEX - (temporal expression)
- Relative position indicators: < and >:
  - *SUBJ> - verb is positioned to the right
  - *SUBJ< - verb is positioned to the left

Examples

- {*SUBJ> [NP vagnstjórinn NP] *SUBJ>} [VP sá VP] {*OBJ< [NP mig NP] *OBJ<} (driver-the saw me)
- {*SUBJ> [NP systir NP] {*QUAL [NP hennar NP] *QUAL} *SUBJ>} [VPb var VPb] (sister her was)
- [VPb er VPb] {*SUBJ< [NP ég NP] *SUBJ<} {*COMP< [VPp fædd VPp] [CP og CP] [VPp uppalin VPp] *COMP<} (am I born and raised)

Efficiency and Error rate

- Written in Java
- Can process 11.300 word-tag pairs per second
  - the output of each module is not written to file, but streamed into the next one
- Accuracy:
  - 96,7% for constituents
  - 84,3% for syntatic functions

Results for various phrase types

Phrase type	F-measure using correct POS tag	F-measure using IceTagger	Freq. in test data
AdvP	91,8%	85,1%	8,2%
AP	95,1%	86,3%	8,1%
APs	87,0%	68,6%	0,5%
NP	96,8%	93,0%	37,6%
NPs	80,4%	74,3%	1,5%
PP	96,7%	91,3%	13,0%
VPx	99,2%	93,8%	19,3%
CP	100%	99,6%	5,7%
SCP	99,6%	97,6%	3,4%
InjP	100%	96,3%	0,2%
MWE	96,9%	92,6%	2,5%
All	96,7%	91,9%	100,0%

Types of errors

- Example of an Adverb phrase error:
  - "um það vissi stúlkan ekki þá" - [PP um [NP það NP] PP] [VP vissi VP] [NP stelpan NP] [AdvP ekki þá AdvP] (about that knew girl not then).
    - "ekki þá" does not belong together
- Errors in adjective phrases:
  - "og tóku fram eigin dósir" - [CP og CP] [VP tóku VP] [NP [AP [AdvP fram AdvP] eigin AP] dósir NP] (and took out own cans)
    - "fram eigin dósir" - "fram" belongs with "tóku"

cont.

- Noun phrase errors:
  - "sterkur var hann og íþróttamaður góður" - [AP sterkur AP] [VPb var VPb] [NPs [NP hann NP] [CP og CP] [NP íþróttamaður NP] NPs] [AP ágætur AP] (strong was he and athlete fine).
  - "hann og íþróttamaður" don't belong together, but are incorrectly parsed as such

Room for improvement?

- Possible options:
  - after shallow parsing, build a deep parse tree
  - use more information in the POS tags - currently only word class and case features are used
    - however, this would mean that the tool could not be used in grammar checking applications