IceParser: An Incremental Finite-State Parser for Icelandic

Hrafn Loftsson[1]   Eiríkur Rögnvaldsson[2]

[1] Department of Computer Science, Reykjavik University, Iceland
[2] Department of Icelandic, University of Iceland, Iceland

NoDaLiDa 2007

Outline

1. Introduction
2. The Icelandic language
3. The annotation scheme
4. IceParser
   - The phrase structure module
   - The syntactic functions module
5. Evaluation
6. Summary

Outline

1. Introduction
2. The Icelandic language
3. The annotation scheme
4. IceParser
   - The phrase structure module
   - The syntactic functions module
5. Evaluation
6. Summary

Finite-state parsing (a form of shallow parsing)

Reductionist approach (Koskenniemi et al., 1992)

- Syntactic tags are associated with words.
- All possible readings of a sentence are reduced to one correct reading using elimination rules.

Constructive approach

- Consists of a collection of syntactic patterns.
- Syntactic labels are inserted into the input strings, e.g.:
  - Brackets denoting constituent structure.
  - Names for grammatical functions.
- A sequence of transducers ⇒ incremental finite-state parsing.
- Xerox Finite-State Tool (XFST) (Karttunen et al., 1996)

Finite-state parsing (a form of shallow parsing)

Reductionist approach (Koskenniemi et al., 1992)

- Syntactic tags are associated with words.
- All possible readings of a sentence are reduced to one correct reading using elimination rules.

Constructive approach

- Consists of a collection of syntactic patterns.
- Syntactic labels are inserted into the input strings, e.g.:
  - Brackets denoting constituent structure.
  - Names for grammatical functions.
- A sequence of transducers ⇒ incremental finite-state parsing.
- Xerox Finite-State Tool (XFST) (Karttunen et al., 1996)

Motivation for developing a finite-state parser

- No parser has been published for Icelandic.
- Shallow parsing is sufficient for many NLP applications, e.g.:
    - Information Extraction
    - Question answering
    - Some types of grammar checking
- Part of the *IceNLP* tool, which itself is a part of a BLARK (Basic Language Resource Kit)
- Efficiency is important ⇒ Finite-state parser

Outline

1. Introduction
2. The Icelandic language
3. The annotation scheme
4. IceParser
   - The phrase structure module
   - The syntactic functions module
5. Evaluation
6. Summary

The Icelandic language

Heavily inflected

- Nouns: three genders, four cases, two numbers, sometimes suffixed definite article.
- Adjectives: four cases, three genders, two numbers, three degrees, “strong” and “weak” form.
- Verbs: three persons, two moods, two tenses, two voices.
- Word order is relatively free.

The POS tagset

- Large, about 660 tags.
- Example: “hestarnir” (horses) ⇒ *nkfng*; noun (n), masculine (k), plural (f), nominative (n), and suffixed definite article (g).

The Icelandic language

Heavily inflected

- Nouns: three genders, four cases, two numbers, sometimes suffixed definite article.
- Adjectives: four cases, three genders, two numbers, three degrees, “strong” and “weak” form.
- Verbs: three persons, two moods, two tenses, two voices.
- Word order is relatively free.

The POS tagset

- Large, about 660 tags.
- Example: “hestarnir” (horses) ⇒ *nkfng*; noun (n), masculine (k), plural (f), nominative (n), and suffixed definite article (g).

Outline

1 Introduction

2 The Icelandic language

3 The annotation scheme

4 IceParser
- The phrase structure module
- The syntactic functions module

5 Evaluation

6 Summary

The annotation scheme (Loftsson & Rögnvaldsson, 2006)

Theory-neutral shallow annotation

- Constituent structure
  - Standard labels: AdvP, AP, NP, PP, VP
  - Additionally: CP, SCP, InjP, MWE, APs, NPs
  - [NP . . . NP], [VP . . . VP]
  - [VPx . . . VPx]; x ϵ {i, b, s, p, g}
- Functional tags
  - Subjects and objects/complements: *SUBJ, *OBJ, *IOBJ, *OBJAP, *OBJNOM, *COMP
  - Other: *QUAL, *TIMEX
  - Relative position indicator, e.g.: *SUBJ>
    (the verb is positioned to the right of the subject)

The annotation scheme (Loftsson & Rögnvaldsson, 2006)

Theory-neutral shallow annotation

- Constituent structure
  - Standard labels: AdvP, AP, NP, PP, VP
  - Additionally: CP, SCP, InjP, MWE, APs, NPs
  - [NP . . . NP], [VP . . . VP]
  - [VPx . . . VPx]; x ϵ {i, b, s, p, g}
- Functional tags
  - Subjects and objects/complements: *SUBJ, *OBJ, *IOBJ, *OBJAP, *OBJNOM, *COMP
  - Other: *QUAL, *TIMEX
  - Relative position indicator, e.g.: *SUBJ>
    (the verb is positioned to the right of the subject)

The annotation scheme

Some 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)

The annotation scheme

Some 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)

The annotation scheme

Some 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)

Outline

1 Introduction

2 The Icelandic language

3 The annotation scheme

4 IceParser

- The phrase structure module
- The syntactic functions module

5 Evaluation

6 Summary

IceParser

Design

- Produces annotations according to our annotation scheme.
- An incremental finite-state parser.
- A purely constructive parser.
- Consists of two modules:
    - The *phrase structure module* (14 transducers).
    - The *syntactic functions module* (8 transducers).

Implementation language

- Java and JFlex (a lexical analyser generator tool); the resulting Java code is a DFA.
- XFST is not used.

IceParser

Design

- Produces annotations according to our annotation scheme.
- An incremental finite-state parser.
- A purely constructive parser.
- Consists of two modules:
  - The *phrase structure module* (14 transducers).
  - The *syntactic functions module* (8 transducers).

Implementation language

- Java and JFlex (a lexical analyser generator tool); the resulting Java code is a DFA.
- XFST is not used.

The transducers

- Include numerous syntactic patterns.
- The actions add syntactic information into the text.
- Rely mainly on word class and subclass information from POS tags.
- The syntactic functions module uses the grammatical *case* feature.

Outline

1. Introduction

2. The Icelandic language

3. The annotation scheme

4. IceParser
   - The phrase structure module
   - The syntactic functions module

5. Evaluation

6. Summary

The phrase structure module

- Adds brackets and labels to indicate constituent structure.
- Input to first transducer is POS tagged text.
- Deepest constituents are analysed first; AdvP ⇒ AP ⇒ NP
- Consider the patterns of the AP transducer:

```
Adj={WordSpaces}{AdjTag}
OpenAdvP="[AdvP"   CloseAdvP="AdvP]"
AdvPhrase={OpenAdvP}~{CloseAdvP}
AdjPhrase={AdvPhrase}?{Adj}
```

- [AdvP mjög aa AdvP] góður lkensf
  (very good)
- [AP [AdvP mjög aa AdvP] góður lkensf AP]

The phrase structure module

- Adds brackets and labels to indicate constituent structure.
- Input to first transducer is POS tagged text.
- Deepest constituents are analysed first; AdvP ⇒ AP ⇒ NP
- Consider the patterns of the AP transducer:

  ```
  Adj={WordSpaces}{AdjTag}
  OpenAdvP="[AdvP"  CloseAdvP="AdvP]"
  AdvPhrase={OpenAdvP}~{CloseAdvP}
  AdjPhrase={AdvPhrase}?{Adj}
  ```

- [AdvP mjög aa AdvP] góður lkensf
  (very good)
- [AP [AdvP mjög aa AdvP] góður lkensf AP]

The phrase structure module

- Adds brackets and labels to indicate constituent structure.
- Input to first transducer is POS tagged text.
- Deepest constituents are analysed first; AdvP ⇒ AP ⇒ NP
- Consider the patterns of the AP transducer:

  ```
  Adj={WordSpaces}{AdjTag}
  OpenAdvP="[AdvP"  CloseAdvP="AdvP]"
  AdvPhrase={OpenAdvP}~{CloseAdvP}
  AdjPhrase={AdvPhrase}?{Adj}
  ```
- [AdvP mjög aa AdvP] góður lkensf
  (very good)
- [AP [AdvP mjög aa AdvP] góður lkensf AP]

The phrase structure module

- Adds brackets and labels to indicate constituent structure.
- Input to first transducer is POS tagged text.
- Deepest constituents are analysed first; AdvP ⇒ AP ⇒ NP
- Consider the patterns of the AP transducer:

  ```
  Adj={WordSpaces}{AdjTag}
  OpenAdvP="[AdvP"   CloseAdvP="AdvP]"
  AdvPhrase={OpenAdvP}~{CloseAdvP}
  AdjPhrase={AdvPhrase}?{Adj}
  ```
- [AdvP mjög aa AdvP] góður lkensf
  (very good)
- [AP [AdvP mjög aa AdvP] góður lkensf AP]

The phrase structure module

- The NP transducer is the most complicated.
- Due to the various ways an NP can be formed.
- The resulting DFA consists of about 50,000 states.
- [AP [AdvP mjög AdvP] góður AP] kennari
  (very good teacher)
- [NP [AP [AdvP mjög AdvP] góður AP] kennari NP]

The phrase structure module

- The NP transducer is the most complicated.
- Due to the various ways an NP can be formed.
- The resulting DFA consists of about 50,000 states.
- [AP [AdvP mjög AdvP] góður AP] kennari
  (very good teacher)
- [NP [AP [AdvP mjög AdvP] góður AP] kennari NP]

The phrase structure module

- The NP transducer is the most complicated.
- Due to the various ways an NP can be formed.
- The resulting DFA consists of about 50,000 states.
- [AP [AdvP mjög AdvP] góður AP] kennari
  (very good teacher)
- [NP [AP [AdvP mjög AdvP] góður AP] kennari NP]

Outline

1 Introduction

2 The Icelandic language

3 The annotation scheme

4 IceParser
- The phrase structure module
- The syntactic functions module

5 Evaluation

6 Summary

The syntactic functions module

- Adds brackets and labels to indicate syntactic functions.
- Input to first transducer: Output of last transducer in the phrase structure module.
- Consider a part of the patterns of the COMP transducer:

```
Compl={APSeqNom}|{NPSeqNom} |
      {VPPastSeq}
SubjVerbBe={Subject}{WS}+{VPBe}{WS}+
SubjVerbCompl={SubjVerbBe}{Compl}
```

The syntactic functions module

- Adds brackets and labels to indicate syntactic functions.
- Input to first transducer: Output of last transducer in the phrase structure module.
- Consider a part of the patterns of the COMP transducer:

```
Compl={APSeqNom}|{NPSeqNom} |
      {VPPastSeq}
SubjVerbBe={Subject}{WS}+{VPBe}{WS}+
SubjVerbCompl={SubjVerbBe}{Compl}
```

The syntactic functions module

An example

- {*SUBJ> [NP hann NP] *SUBJ>} [VPb er VPb] [NP [AP [AdvP mjög AdvP] góður AP] kennari NP]
- (he is (a) very good teacher)
- {*SUBJ> [NP hann NP] *SUBJ>} [VPb er VPb] [NP [AP {*COMP< [AdvP mjög AdvP] góður AP] kennari NP] *COMP<}

The syntactic functions module

An example

- {*SUBJ> [NP hann NP] *SUBJ>} [VPb er VPb] [NP [AP [AdvP mjög AdvP] góður AP] kennari NP]
- (he is (a) very good teacher)
- {*SUBJ> [NP hann NP] *SUBJ>} [VPb er VPb] [NP [AP {*COMP< [AdvP mjög AdvP] góður AP] kennari NP] *COMP<}

Outline

1. Introduction
2. The Icelandic language
3. The annotation scheme
4. IceParser
   - The phrase structure module
   - The syntactic functions module
5. Evaluation
6. Summary

Evaluation

Experimental setup

- A *gold standard* was constructed:
  - About 500 sentences randomly selected from the POS tagged *IFD* corpus.
  - Manually annotated with constituent structure and syntactic functions using the annotation scheme.
- The *Evalb* (Sekine & Collins, 1997) bracket scoring program used for automatic evaluation.
- The parser evaluated using correct POS tags and tags generated by *IceTagger* (Loftsson, 2006).
  - POS tagging accuracy was 91.1% (unknown word ratio 7.8%).

Results for the various phrase types

Phrase type	F-measure using correct POS tags	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	100.0%	99.6%	5.7%
SCP	99.6%	97.6%	3.4%
InjP	100.0%	96.3%	0.2%
MWE	96.9%	92.6%	2.5%
All	96.7%	91.9%	100.0%

Constituents: A comparison

- First parser evaluation published for Icelandic.
- Comparison with Swedish:

	F-measure		
Parser	All phrases	NP	Tagger
IceParser	96.7%	96.8%	No
Kokkinakis & J.-Kokkinakis (1999)	93.3%	96.2%	Yes (98.7%)
IceParser	91.9%	93.0%	Yes (91.1%)
Knutsson et al. (2003)*	88.7%	91.4%	Yes

* not finite-state

Results for the various syntactic functions

Function type	F-measure using correct POS tags	F-measure using *IceTagger*	Freq. in test data
SUBJ	68.2%	47.6%	4.7%
SUBJ>	92.7%	89.4%	30.3%
SUBJ<	83.7%	75.1%	12.3%
OBJ	0.0%	0.0%	0.2%
OBJ>	43.5%	20.0%	0.8%
OBJ<	90.2%	78.2%	19.7%
OBJAP>	71.4%	57.2%	0.2%
OBJAP<	75.0%	46.2%	0.4%
OBJNOM<	30.8%	16.7%	0.6%
...			
All	84.3%	75.3%	100.0%

Syntactic functions: A comparison

- Comparison with German:

	F-measure			
Parser	All functions	SUBJ	OBJ	Tagger
IceParser	84.3%	90.5%	88.2%	No
Müller (2004)	82.5%	90.8%	64.5% (acc.) 81.9% (dat.)	No

Efficiency

Method	Word-tag pairs per sec.	Speed increase
Writing output to files	6,700	
Writing output to memory	11,300	75%

Outline

1. Introduction
2. The Icelandic language
3. The annotation scheme
4. IceParser
   - The phrase structure module
   - The syntactic functions module
5. Evaluation
6. Summary

Summary

- *IceParser* is an incremental finite-state parser, based on a shallow annotation scheme.
  - A phrase structure module.
  - A syntactic functions module.
- *IceParser* is both effective and efficient.
- Future work:
  - Improve individual components.
  - Build a version which uses the morphological info in POS tags to a greater extent.
- The parser can be tested by visiting http://nlp.ru.is