T-(538|725)-MALV, Natural Language Processing
Partial parsing – IceParser

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]School of Computer Science, Reykjavik University

October 2009

Outline

Outline

Incremental finite-state parsers

(e. cascading/incremental partial/shallow/finite-state parsers)

- Based on a sequence of finite-state transducers (í. stöðuferjöldum)
- Each transducer has a specific role, for example to:
  - Mark MWEs
  - Mark verb phrases
  - Mark noun phrases
  - Mark preposition phrases
  - etc.
- The input is a tokenised and POS tagged text.
- The output of one transducer is used as the input to the next transducer in the sequence.
- The final output is the original text marked with syntactic information (e.g. chunks).

Incremental finite-state parsers

Exist for various languages

- Spanish (Molina et al. 1999)
- Swedish (Megyesi and Rydin 1999)
- German (Müller 2004)
- French (Aït-Mokhtar and Chanod 1997)
- Icelandic (Hrafn Loftsson and Eiríkur Rögnvaldsson 2007)

Efficient

- A sequence of finite-state transducers.

Annotation scheme

Purpose

- Description for annotation of constituent structure and syntactic functions.
- Grammar definition corpus (GDC) (í. málfræðiskilgreiningarmálheild).
  - A collection of typical sentences which have been annotated according to the annotation scheme.
  - Its purpose is to "provide an unambiguous answer to the question how to analyse any utterance in the object language" (Voutilainen, 1997).
  - Furthermore, a GDC can be used to develop a parser, because a parser should at least be able to analyse correctly the sentences in the GDC.

Outline

IceParser

- Hrafn Loftsson and Eiríkur Rögnvaldsson, 2007
- Based on an annotation scheme:
  `http://nlp.ru.is/pdf/shallowAnnotation.pdf`
- An incremental finite-state parser: `http://nlp.ru.is`
- Annotates constituents and syntactic functions.
  - Phrase/constituent structure module; 14 transducers.
  - Syntactic functions module; 8 transducers.

How is the accuracy estimated?

Constituents	Correct constituents in *gold standard*	Incorrect constituents
Generated by parser	A	B
Not generated by parser	C	

- **Precision**: $P = \frac{A}{A \cup B} = \frac{\#\text{ correct constituents in output of parser}}{\text{total }\#\text{ of constituents in output of parser}}$
- **Recall**: $R = \frac{A}{A \cup C} = \frac{\#\text{ correct constituents in output of parser}}{\text{total }\#\text{ of constituents in } \textit{gold standard}}$
- **F-measure** $= \frac{2 * P * R}{P + R}$ (e. harmonic mean)

IceParser: 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*.
  - 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%

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%