Statistical Parsing of Morphologically Rich Languages How, Where and Whither

SPMRL'10 – A Gentle Introduction

Reut Tsarfaty, Uppsala Universitet
Djamé Seddah, Alpage (Inria/Univ. Paris-Sorbonne)
Yoav Goldberg, Ben Gurion University
Sandra Kübler, Indiana University
Marie Candito, Alpage (Inria/Univ. Paris 7)
Jennifer Foster, NCLT, Dublin City University
Yannick Versley, Universität Tübingen
Ines Rehbein, Universität Saarbrücken
Lamia Tounsi, NCLT, Dublin City University

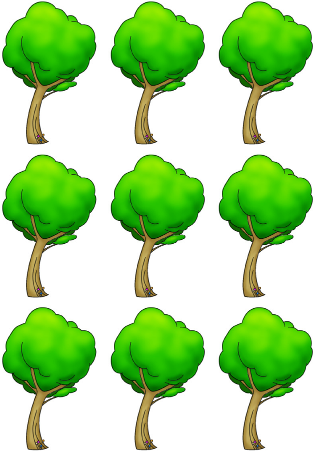
Statistical Parsing

Statistical Parsing

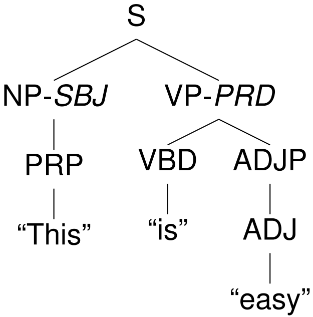
“This is easy”

Statistical Parsing

“This is easy”



Statistical Parsing


S

NP-*SBJ*

VP-*PRD*

PRP

VBD

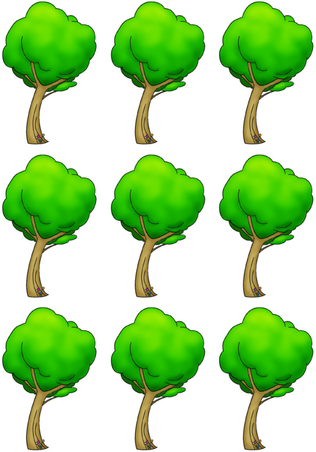
ADJP

“This”

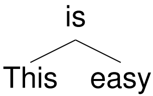
“is”

ADJ

“easy”

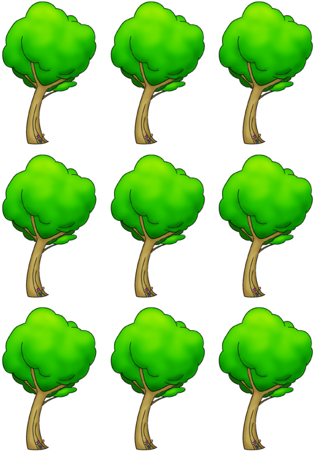


Statistical Parsing



is

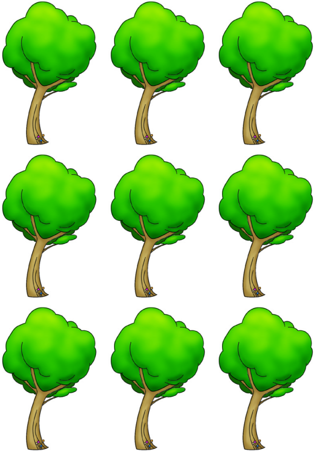
This easy



Statistical Parsing

[This [is [easy]]]

Supervised Statistical Parsing



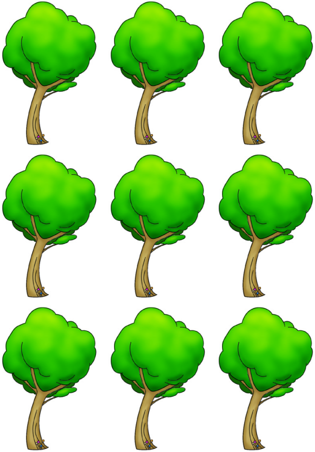
Supervised Statistical Parsing

Constituency-Based:

Accuracy results for German, French, Korean, Arabic, Hebrew and others lag behind those for English.

Dependency-Based:

CoNLL Shared Task: Arabic, Basque and Greek show the lowest performance regardless of the parser used.



So What Is Going On?

Often Considered..

- Corpora Size
  E.g., For *Chinese* (Bikel & Chiang 2000)
- Annotation Idiosyncrasies
  E.g., For *Arabic* (Maamouri, Bies & Kulick 2008, 2009)
- Evaluation Matters
  E.g., For *German* (Rehiben & van Genabith 2007, Kübler 2008)

So What Is Going On?

Often Considered..

- Corpora Size
  E.g., For *Chinese* (Bikel & Chiang 2000)
- Annotation Idiosyncrasies
  E.g., For *Arabic* (Maamouri, Bies & Kulick 2008, 2009)
- Evaluation Matters
  E.g., For *German* (Rehiben & van Genabith 2007, Kübler 2008)

A Recurring Trend

English, Chinese $>$ German, French $>$ Hebrew, Arabic

Defining Morphologically Rich languages (MRLs)

Morphology

High Synthesis (high morpheme/words ratio)

High Fusion (non-concatenative morphology)

Defining Morphologically Rich languages (MRLs)

Morphology

High Synthesis (high morpheme/words ratio)

High Fusion (non-concatenative morphology)

Syntax

Free Word-Order

Discontinuous Constituents

Null Elements

Defining Morphologically Rich languages (MRLs)

Morphology

High Synthesis (high morpheme/words ratio)

High Fusion (non-concatenative morphology)

Syntax

Free Word-Order

Discontinuous Constituents

Null Elements

Morphosyntax

Case/government

Agreement

Clitics

Parsing with MRLs: Shared Challenges

Architectural Aspects

Modeling Aspects

Learning Aspects

Parsing with MRLs: Shared Challenges

Architectural Aspects

- What is the input? Words? Morphemes?
- If Words – Which abstract representation?
- If Morphemes – When to morphologically analyze?

Modeling Aspects

Learning Aspects

Parsing with MRLs: Shared Challenges

Architectural Aspects

- What is the input? Words? Morphemes?
- If Words – Which abstract representation?
- If Morphemes – When to morphologically analyze?

Modeling Aspects

- What morphological information?
- What morphosyntactic representation?
- How to deal with nonconfigurational structures?

Learning Aspects

Parsing with MRLs: Shared Challenges

Architectural Aspects

- What is the input? Words? Morphemes?
- If Words – Which abstract representation?
- If Morphemes – When to morphologically analyze?

Modeling Aspects

- What morphological information?
- What morphosyntactic representation?
- How to deal with nonconfigurational structures?

Learning Aspects

- How to deal with lexical sparsity?
- How to deal with syntactic sparsity?
- How to deal with bi-lexical dependencies?

Today:

Session: Dependency Parsing of MRLs

- Arabic Dependency Parsing with Lexical/Morphological Features
- Local Morphosyntactic Features in Hindi Dependency Parsing
- Different Techniques for Dependency Parsing of Basque

Session: Constituency Parsing of MRLs

- Modeling Agreement for Modern Hebrew Parsing
- Factors Affecting the Accuracy of Korean Parsing
- Direct Parsing of Discontinuous Constituents

Session: Estimation and Lemmatization

- Unknown words in LA parsing for English, Arabic, French
- Parsing Word Clusters (for French)
- Lemmatization and Lexicalized Parsing for French

Today (cont.):

Session: Dependency Parsing of MRLs

- Morphosyntactic Features in Hindi Dependency Parsing
- Easy-First Hebrew Depedency Parsing

Invited Talk by Kevin Knight

- Morphology in Statistical Machine Translation

Panel Discussion

- Dan Bikel
- Julia Hockenmaier
- Slav Petrov
- Owen Rambow

Today (cont.):

From a Bird's Eye View

	Constituency-Based	Dependency-Based
Arabic	X	X
Basque	-	X
English	X	-
French	XXX	-
German	X	-
Hebrew	X	X
Hindi	-	XX
Korean	X	-

Table: An overview of SPMRL contributions.

Overarching Questions

- Evaluation
  - Across Languages
  - Across Treebanks
  - Across Frameworks

Overarching Questions

- Evaluation
  - Across Languages
  - Across Treebanks
  - Across Frameworks
- Annotation
  - Universality
  - Diversity
  - Interpretability

Overarching Questions

- Evaluation
  - Across Languages
  - Across Treebanks
  - Across Frameworks
- Annotation
  - Universality
  - Diversity
  - Interpretability
- Applications
  - Statistical Machine Translation

Workshop Goals

- To increase visibility
- To identify recurring problems
- To discuss shared solutions

(See also: overview paper in the proceedings)

So, Sit Back and Relax...

Enjoy The Ride!!