

ELSNET Roadmap for CL

Area Report: Parsing

John Carroll

University of Sussex, UK

Broad view:

- POS tagging, morphology
- assigning structure
- grammar formalisms
- data obtained through / used in parsing

WSD in Agirre *et al.* presentation

Conference (and Workshop) Topics

Word / sentence segmentation

Morphology Arabic / Asian languages

Chunking syntactic / topological / prosodic groups

Annotating corpora

POS tag correction, annotation schemes and procedures,
enriching / converting treebanks

Parsing sub-tasks

PP attachment, Japanese zero-anaphor detection / case
assignment, Korean adnoun clause structure

Grammar formalisms / formal properties

modelling word order, grammar representation, learn-
ability, parsing termination, development strategies

Parsing and speech recognition

structured language models, robust parsing of speech

Integration

deep and shallow, general and domain-specific

Unification grammar

probabilistic models, dependency parsing, defaults

Statistical parsing

inducing grammars, combining features / models,
lexicalisation

Lexical acquisition

word clustering, subcategorisation, semantic roles

Semantics

syntax / semantics interface, efficient interpretation

Algorithmic frameworks for parsing

Current Trends

Recent breakthroughs:

- efficient deep processing
- non-English treebanks
- lexicalised probabilistic models

More:

- deep analysis, integration of deep and shallow
- work in languages other than English — and tackling phenomena not occurring in English

Less (at least at this conference):

- parsing algorithms

The Future

Common resources and accepted evaluation procedures

- processing tools — both generic and pre-packaged (TTT, TnT, Abney, Collins, Link Grammar, Minipar, ...)
- generic data (COMLEX, WordNet, Levin classification, ...)
- common data for comparative parser evaluation

Annotated corpora / treebanks

- in more languages, containing more detail / semantics
- methods for semi-automated annotation, checking

Unsupervised lexical acquisition

- manual development of lexical resources ranges from expensive to impractical

Application / domain-specific tuning of off-the-shelf components

- integrating results of lexical acquisition
- combining functionally similar modules (e.g. running alternative components in parallel)
- machine learning for adaptation and combination of results

More deep analysis and manual grammar development

- suitable formalisms exist (HPSG, LFG, LTAG, CG) — but enough skilled grammarians?
- probabilistic models for deep grammars