Intrinsic and Extrinsic Evaluations

[Galliers and Sparck Jones (1993)]

- Intrinsic Evaluation Criteria:
  - Relating to a system's objective
- Extrinsic Evaluation Criteria:
  - Relating to the system's function i.e. to its role in relation to its setup's purpose

CENTRE FOR LANGUAGE TECHNOLOGY
MACQUARIE UNIVERSITY - SYDNEY

Intrinsic versus Extrinsic Evaluations of Parsing Systems

Diego Mollá-Aliod
Ben Hutchinson

14 April 2003

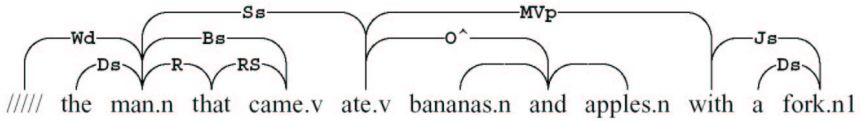
- Link Grammar and Conexor FDG
- Intrinsic Evaluation
- Extrinsic Evaluation: Answer Extraction
- Discussion

Link Grammar and Conexor FDG

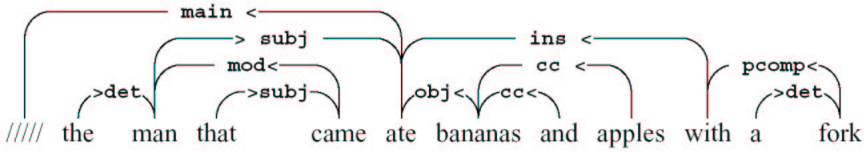
- Two examples of state-of-the-art parsing systems
- Robust treatment of "difficult" or "ungrammatical" sentences
- Dependency-based
- Link Grammar:
  - Publicly available
  - Developed by Carnegie Mellon University
  - http://www.link.cs.cmu.edu/link/
- Conexor Functional Dependency Grammar (Conexor FDG):
  - Proprietary
  - Initially developed by the University of Helsinki
  - http://www.conexor.fi/

Link Grammar and Conexor FDG

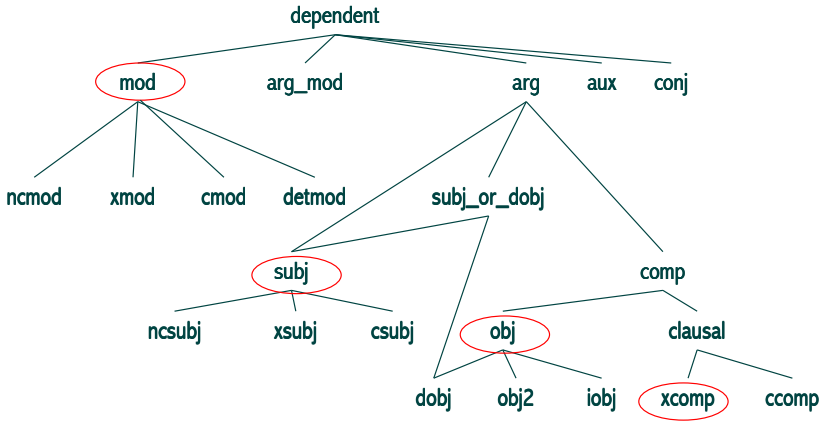
- Link Grammar



- Conexor FDG



Intrinsic Evaluation – Grammatical Relations



Grammatical Relations

- *The man that came ate bananas and apples with a fork.*

  (detmod _ man the) (cmod that man come) (ncsubj come man _) (ncsubj eat man _) (dobj eat banana _) (dobj eat apple _) (conj and banana apple) (ncmod fork eat with) (detmod _ fork a)

- Same example with the selected gramrels

  (mod that man come) (subj come man _) (subj eat man _) (obj eat banana _) (obj eat apple _)
  (mod fork eat with)

Grammatical Relations

- *Failure to do this will continue to place a disproportionate burden on Fulton taxpayers.*

  (xcomp to failure do) (dobj do this _) (ncsubj continue failure _) (xcomp to continue place) (ncsubj place failure _) (dobj place burden _) (ncmod _ burden disproportionate) (iobj on place tax-payer) (ncmod _ tax-payer Fulton) (detmod _ burden a) (aux _ continue will)

- Same example with the selected gramrels

  (xcomp to failure do) (obj do this _) (subj continue failure _) (xcomp to continue place) (subj place failure _) (obj place burden _) (mod _ burden disproportionate) (obj on place tax-payer) (mod _ tax-payer Fulton)

Intrinsic Evaluation

- Corpus [Briscoe & Carroll (2000)]
  - 500 sentences / 10,000 words
  - Annotated with the grammatical relations
- For each grammatical relation type we compute:
  - Precision:

$$\frac{|\text{Correct occurrences retrieved}|}{|\text{Retrieved occurrences}|}$$

  - Recall:

$$\frac{|\text{Correct occurrences retrieved}|}{|\text{Correct occurrences}|}$$

Results of Intrinsic Evaluation

		With Link Grammar	*With Conexor FDG*
Precision	SUBJ	50.3%	73.6%
	OBJ	48.5%	84.8%
	XCOMP	62.2%	76.2%
	MOD	57.2%	63.7%
	Average	*54.6%*	*74.6%*
Recall	SUBJ	39.1%	64.5%
	OBJ	50%	53.4%
	XCOMP	32.1%	64.7%
	MOD	53.7%	56.2%
	Average	*43.7%*	*59.7%*

Extrinsic Evaluation

- Embedding setup: Answer Extraction
  - *Locate those exact phrases of unedited text documents that answer a query worded in natural language*
- ExtrAns
  - An answer extraction system
  - Uses logical forms to determine the answer of a question
  - The version for the present evaluation uses:
    - a full parser (Link Grammar or Conexor FDG);
    - a semantic interpreter;
    - a simple thesaurus based on WordNet;
    - an answer extraction module that operates on logical forms.

The Logical Forms

- Called Minimal Logical forms because they encode the minimum information required for AE
- Flat expressions that use reification

- Example: *cp will quickly copy files*

  holds(e4), object(cp,o1,[x1]), object(s_command,o2,[x1]), evt(s_copy,e4,[x1,x6]), object(s_file,o3,[x6]), prop(quickly,p3,[e4]).
- Example: *the man that came ate bananas and apples with a fork*

  holds(e1), object(s_man,o2,[x2]), evt(s_come,e4,[x2]), evt(s_eat,e5,[x7]), e6@<e7, e8@<e7, evt(s_eat,e5_1,[x6]), evt(s_eat,e5_2,[x8]), object(s_banana,o6,[x6]), object(s_apple,o8,[x8]), prop(with,p9,[e6]), object(s_fork,o11,[x11]).

Scoring the Overlap of Logical Forms

- Synonym mode:
  - Find the synonym representatives
  - Use Prolog resolution
  - Only finds exact matches
- Approximate mode:
  - Find the synonym representatives
  - Compute the highest overlap possible with variable unification
  - Return the sentence(s) with highest overlap
- If there are exact matches, Synonym mode and Approximate mode return the same answers

Extrinsic Evaluation

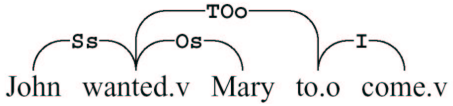
- Corpus:
  - 500 Unix manual pages
  - 26 sample questions with the answers found in the corpus
- Nature of the questions:
  - There is at least one answer in the corpus
  - The question asks how to perform a particular action, or how a particular command works
  - The question is simple
- Precision and recall as in standard Information Retrieval
- F-score = 2 (|Returned and relevant|) / (|Returned| + |Relevant|)

Results of Extrinsic Evaluation

Synonym mode			
Parser	*Precision*	*Recall*	*F-score*
Conexor FDG	55.8%	8.9%	0.074
LG-best	49.7%	11.4%	0.099
LG-all	50.9%	13.1%	0.120
Approximate mode			
Parser	*Precision*	*Recall*	*F-score*
Conexor FDG	28.3%	21.9%	0.177
LG-best	31.8%	15.8%	0.150
LG-all	40.5%	20.5%	0.183

Final Discussion: Intrinsic or Extrinsic Evaluations?

- A "good" parser is not necessarily best for an application?
  - The conversion to grammatical relations may throw away important information
  - Consistent errors/idiosyncrasies in the parser output can be corrected in subsequent processing stages



- Variables introduced in the evaluation may affect the results…

To Do

- Extrinsic evaluation where AE is based on the overlap of grammatical relations
  - To remove variables in the experiments
- Use same corpus for both intrinsic and extrinsic evaluations
  - Any suggestions?
- Intrinsic evaluation of parser+semantic interpreter
- Use other intrinsic evaluations of parsers (e.g. constituency-based)
- Use other embedding setups for extrinsic evaluations
  - To test if similar results occur