Evaluating deep syntactic parse accuracy and its effect in *why*-question answering

This presentation

1. The TOSCA system;

2. Intrinsic parser evaluation;

3. Extrinsic parser evaluation:
   Answer type determination in *why*-question analysis

1. The TOSCA parser

- Deep syntactic parser
  - Categories
  - Features
  - Functions
- Needs human interference at two stages:
  - After tagging → checking tags
  - After parsing → selecting desired tree
- Interactive character problematic in *why*-QA application

1.1 Data

- 238 questions formulated to newspaper texts by native speakers

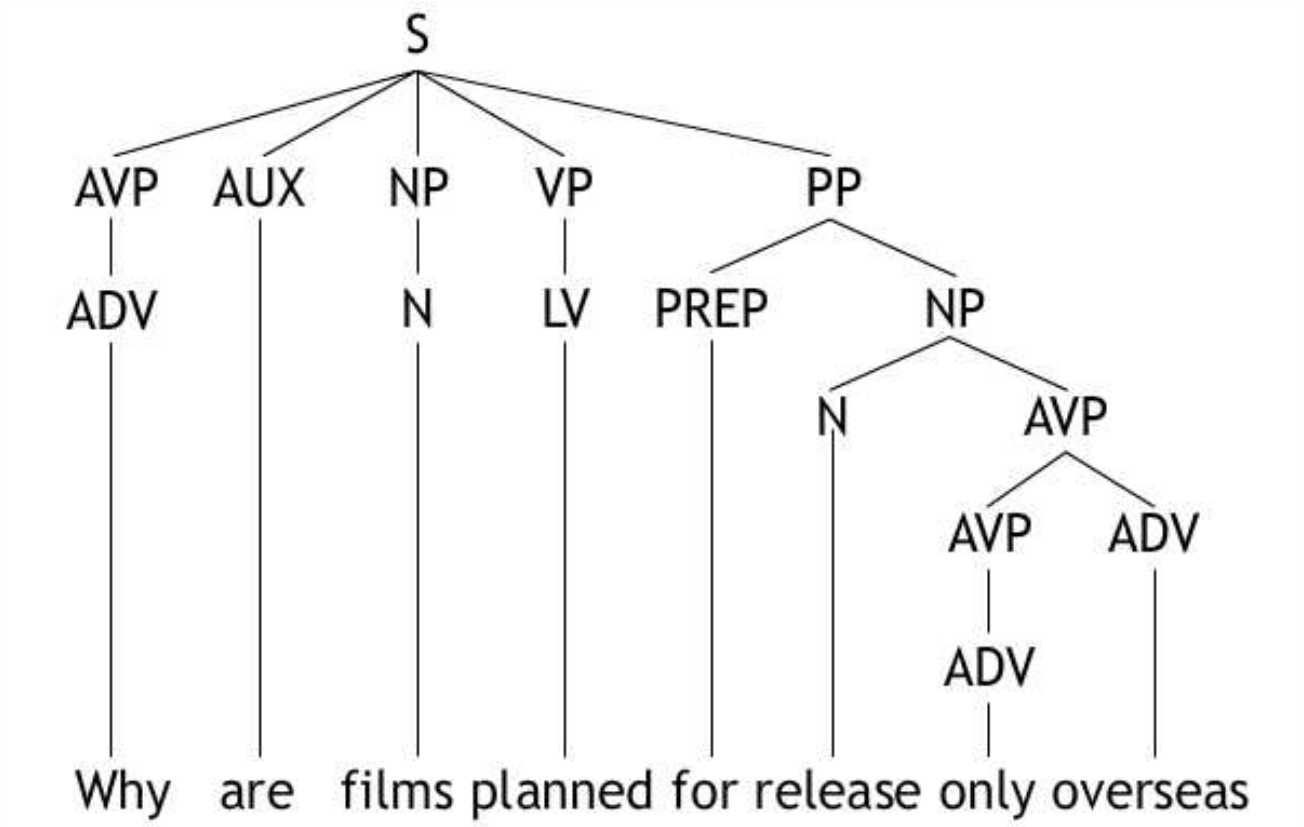
- Three different data sets of questions:

	number	coverage
1. Gold standard	238	100%
2. Semi-automatic output	233	98%
3. Fully automatic output	190	80%

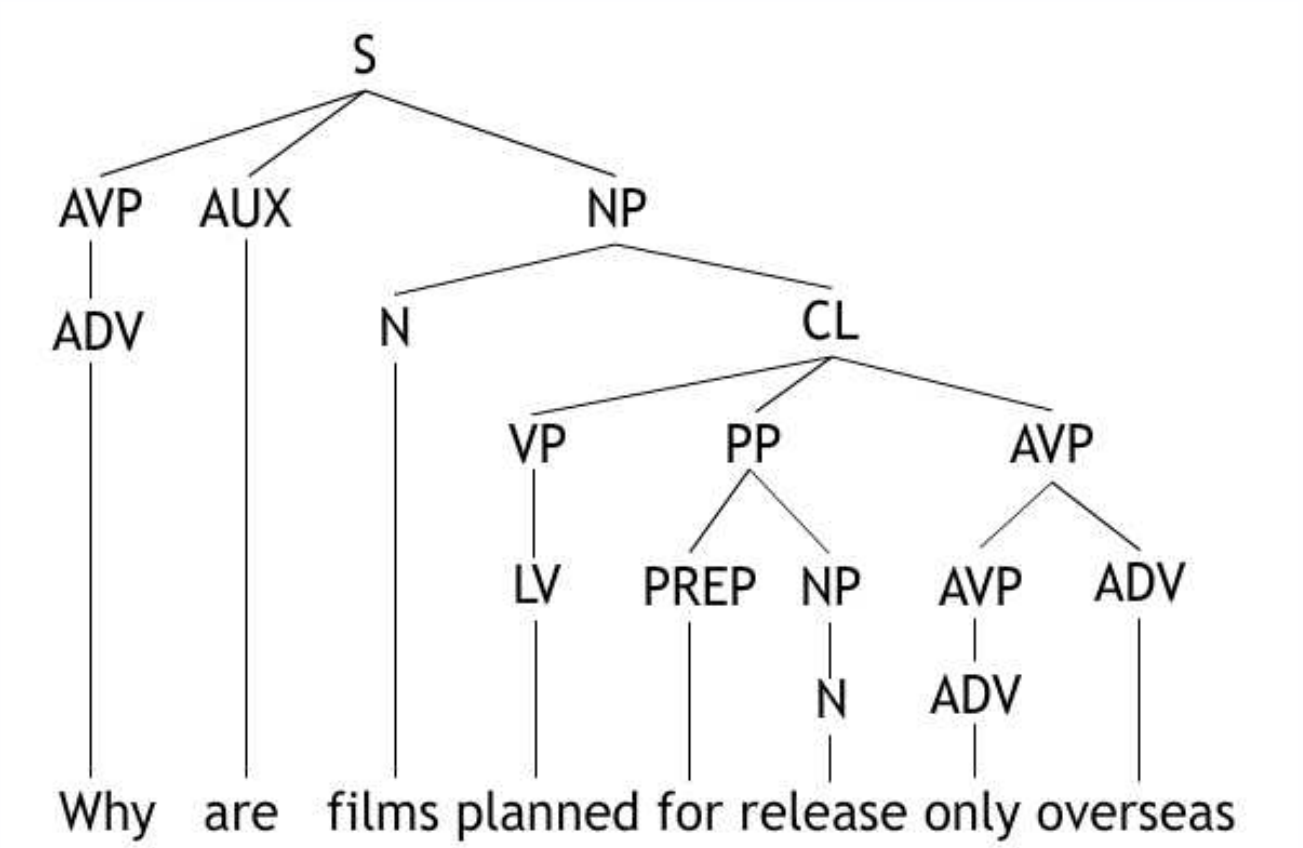
2. Intrinsic Parser Evaluation

- Evaluating parser: use semi-automatic output
- Leaf-Ancestor (LA) assessment
  (Sampson et al., 1989; Sampson, 2000)
- An example:
  - "Why are films planned for release only overseas?"

2.1 Calculating LA scores (1)



2.1 Calculating LA scores (1)



2.1 Calculating LA scores (2)

- Example: “planned”
  Gold standard: LV VP S
  TOSCA output: LV VP [CL NP S

- Total number of labels: 9
  Number of correct labels = 6
  LA score = 6/9 ≈ **0.67**

2.2 Results

- Number of incorrectly analysed questions:

	number:	frequency:
semi-automatic output	**45**	**19.3%**

- Average LA score:

	with error(s):	total:
semi-automatic output	**0.78**	**0.98**

2.3 Parse error analysis (1)

- Common parsing errors (semi-automatic):
  - Incorrect functions:
    - Subject
    - Direct object
    - Subject complement
    - Adverbial
  - Incorrect features, e.g.:
    - Tense
    - Transitivity
    - Modality
  - Incorrect tree structure

2.3 Parse error analysis (2)

- Problemetic words (semi-automatic):

word	Freq.	% with error	word	Freq.	% with error
than	5	60.0%	warming	6	33.3%
chefs	5	60.0%	rights	6	33.3%
for	15	46.7%	global	6	33.3%
court	9	44.4%	at	10	30.0%
supreme	7	42.9%	women	7	28.6%
easier	5	40.0%	and	7	28.6%
dictionary	5	40.0%	up	8	25.0%
with	8	37.5%	in	40	25.0%
about	9	33.3%			

3. Extrinsic parser evaluation

- Influence of parser accuracy: semi-automatic output
- Contribution to *why*-QA application: fully automatic output

- *Why*-QA task: answer type determination

  Answer type for *why*-questions: 'reason'

  'cause' 'circumstance'

  'motivation' 'purpose'

3.1 Determining answer types

- Importance of answer type determination for quality of QA system (Hovy et al., 2001)
- Importance of using syntax for answer type determination (Verberne et al., 2006)

- Values of lexical and syntactic features offered to machine learning algoritms
- Predict answer type

3.2 Results

- Semi-automatic output:

	gold standard Naïve Bayes	semi-automatic Naïve Bayes
233 questions	79.2%	79.2%

- Fully automatic output:
  - 48 questions could not be parsed
  - Guess feature values by using tagger output

	gold standard Naïve Bayes	fully automatic C4.5
238 questions	78.5%	66.3%

Conclusion

- Parser performance excellent, also in *why*-QA task
- Tagger quality hampers use of TOSCA system in *why*-QA

- Questions?