# Models of Synchronous Grammar Induction for SMT

**Phil Blunsom**[1] **Joy Ying Zhang**[2]

[1]University of Oxford
[2]CMU

December 6, 2009

# Workshop overview

Input:

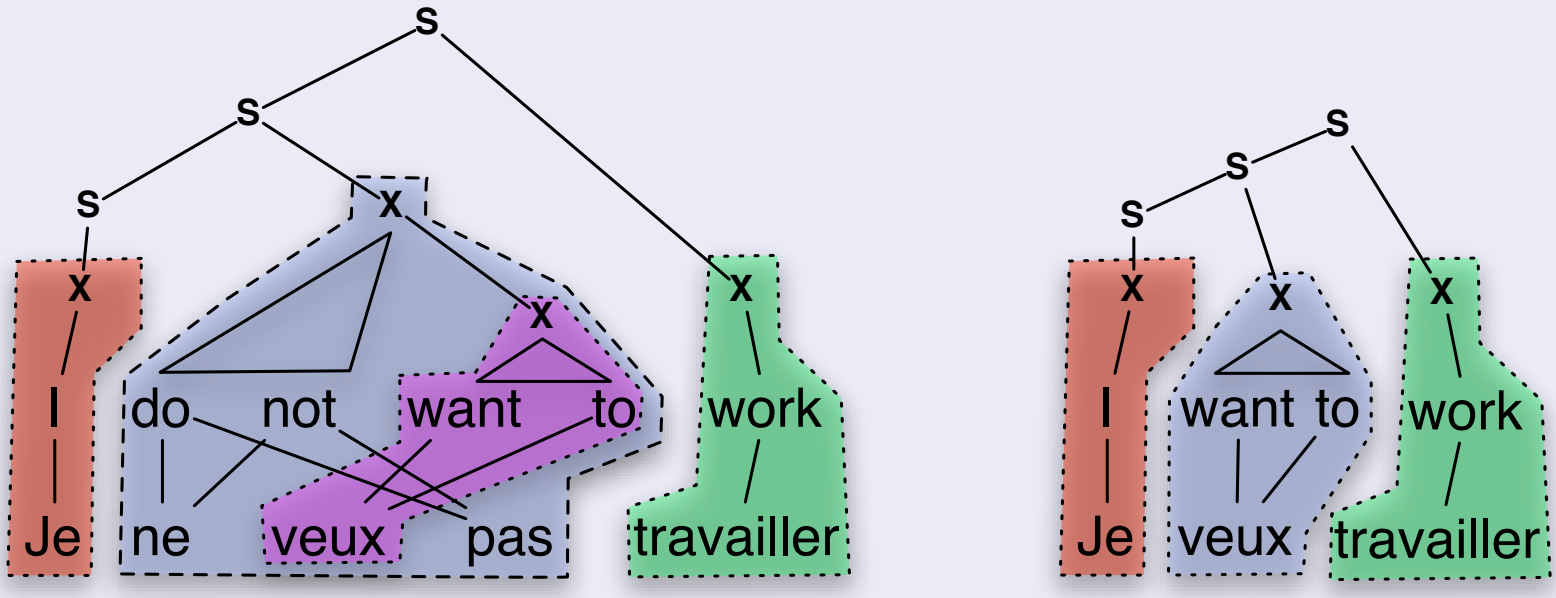
- Existing procedures for synchronous grammar extraction

Output:

- New unsupervised models for large scale synchronous grammar extraction,
- A systematic comparison and analysis of the existing and proposed models.
- An extended Joshua decoder capable of working with these models,

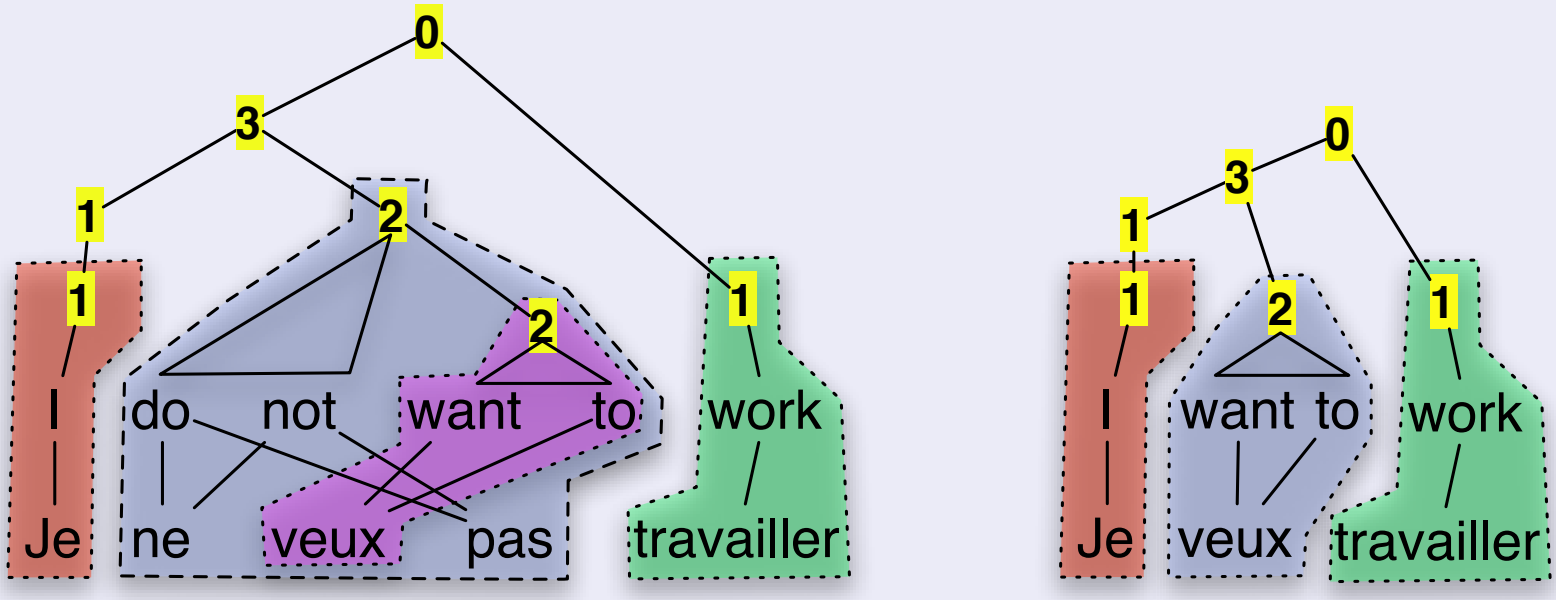
# Models of translation

## Hierarchical



# Models of translation

## Hierarchical



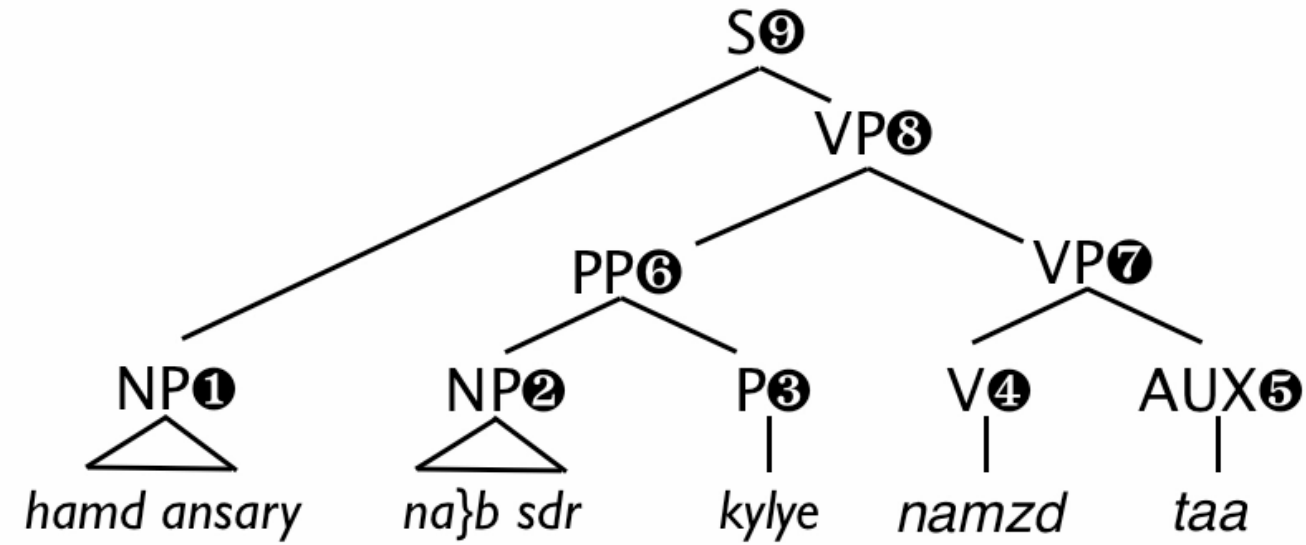

- AIM: Implement a large scale open-source synchronous constituent learning system.
- AIM: Investigate and understand the relationship between the choice of synchronous grammar and SMT performance.

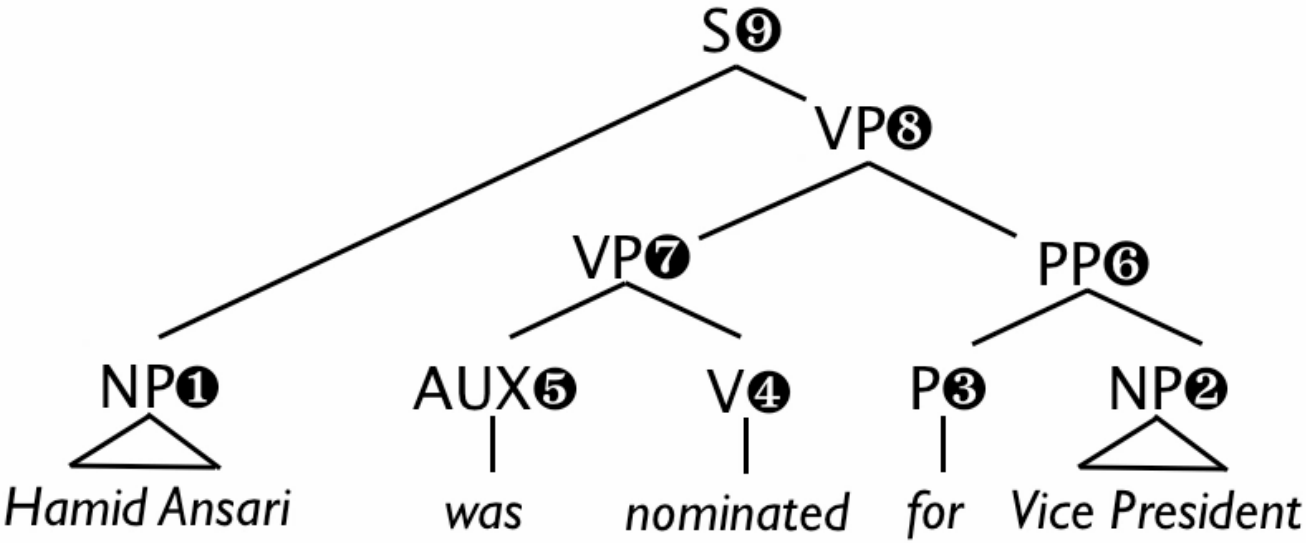
# Impact

Success will have a significant impact on two areas of CL:

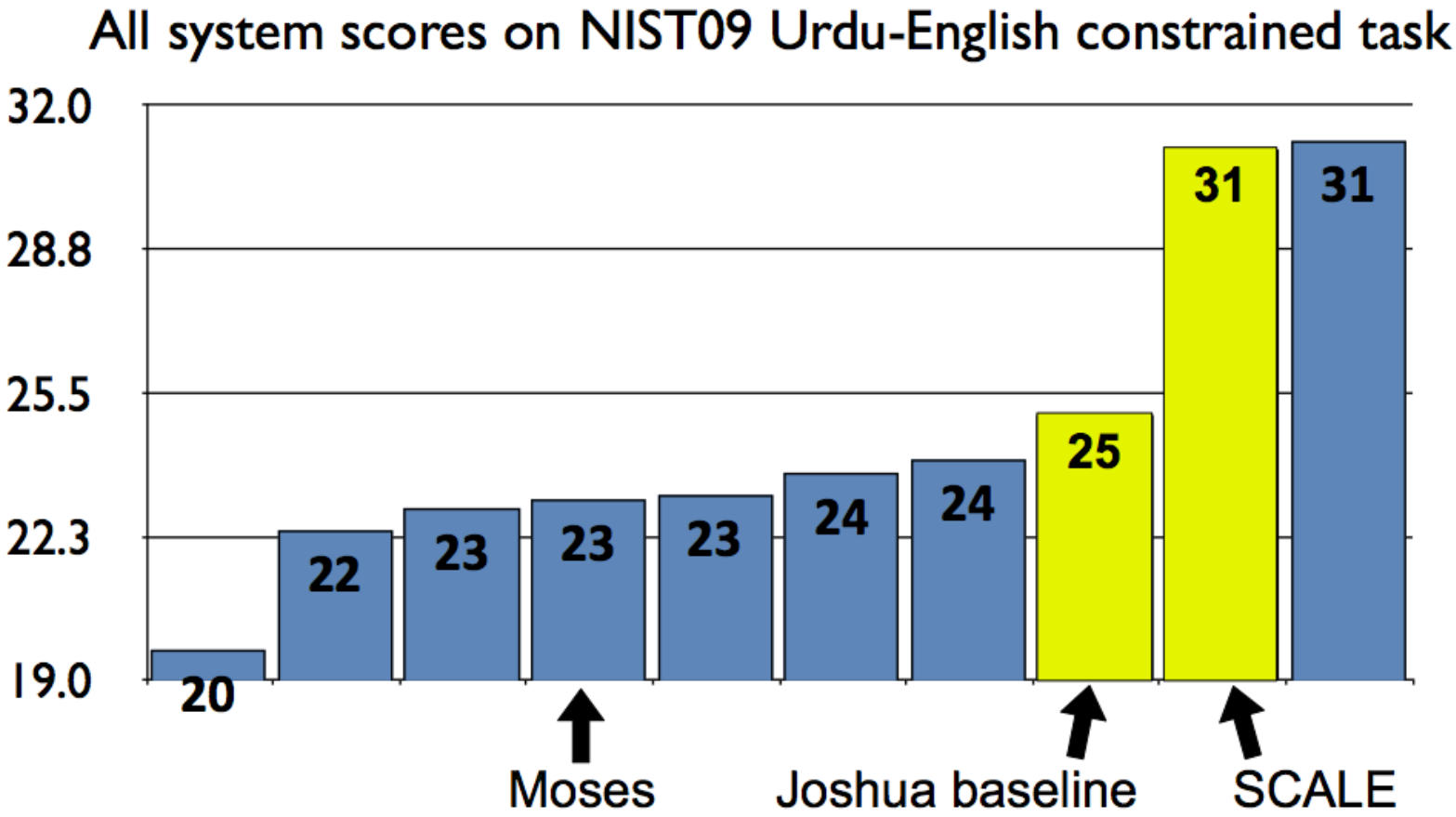
- Grammar induction:
  - Provide an empirical validation of state-of-the-art grammar induction techniques.
- Machine translation
  - Make the benefits of richly structured translation models available to a much wider range of researchers and for a wider range of languages.
  - Change the research outlook of the field.

# Impact


S❾
VP❽
PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa


S❾
VP❽
VP❼
PP❻
NP❶
AUX❺
V❹
P❸
NP❷
Hamid Ansari
was
nominated
for
Vice President

# Impact


All system scores on NIST09 Urdu-English constrained task
32.0
28.8
25.5
22.3
19.0
20
22
23
23
23
24
24
25
31
31
Moses
Joshua baseline
SCALE

# Impact

| Language | Words | Domain |
|---|---|---|
| English | 4.5M | Financial news |
| Chinese | 0.5M | Broadcasting news |
| Arabic | 300K (1M planned) | News |
| Korean | 54K | Military |

Table: Major treebanks: data size and domain

# Impact

# Evaluation goals

Will we predominately evaluate using BLEU, but also use automatic structured metrics and perform small scale human evaluation:

- Evaluate phrasal, syntactic, unsupervised syntactic,
- Aim 1: Do no harm (not true of existing syntactic approach)
- Aim 2: Exceed the performance of current non-syntactic systems.
- Aim 3: Meet or exceed performance of existing syntactic systems.

# Constituency and context


- Design and apply large scale scale clustering and topic modelling algorithms (LDA, HDPs, HPYPs etc),
- identify sets of frequent contexts that distinguish synchronous constituent properties.
- Motivated by successful models of monolingual grammar induction,
- deep connections to unsupervised word sense disambiguation, thesaurus extraction etc.

# Constituency and context


- Design and apply large scale scale clustering and topic modelling algorithms (LDA, HDPs, HPYPs etc),
- identify sets of frequent contexts that distinguish synchronous constituent properties.
- Motivated by successful models of monolingual grammar induction,
- deep connections to unsupervised word sense disambiguation, thesaurus extraction etc.

# Constituency and context


- Design and apply large scale scale clustering and topic modelling algorithms (LDA, HDPs, HPYPs etc),
- identify sets of frequent contexts that distinguish synchronous constituent properties.
- Motivated by successful models of monolingual grammar induction,
- deep connections to unsupervised word sense disambiguation, thesaurus extraction etc.

# Schedule

- Pre-workshop:
  - Collect existing opensource tools for synchronous grammar induction,
  - Collect corpora across a range of tranlations conditions: small, large, low-density languages etc.
  - Design the integtration of various existing approaches into the Joshua decoder.
- Week 1:
  - Optimise and reconfigure decoder to handle labelled synchronous grammars,
  - Perform a empirical study of synchronous constituency models,
  - Implement phrase and context extraction algorithms.

# Schedule

- Week 2-3:
  - Continue optimising decoder to handle labelled synchronous grammars,
  - Implement unsupervised label induction algorithms, initially inducing a single label per-phrase.
  - Extend to "topic"-modelling style representation where a phrase may have multiple labellings.
  - Perform experimental comparison of existing synchronous grammar translation models.
- Week 3-6:
  - Perform experimental comparison of unsupervised synchronous grammar translation models.
  - Extend the evaluation to small/big data sets, hi-density vs. low-density language pairs.
  - Create "semi-supervised" models combining knowledge from treebank parser into the unsupervised algorithms.
  - Wrap-up and write final report.

# Potential team members

**Faculty**

Phil Blunsom (Machine learning, SMT, grammar induction)
Joy Ying Zhang (SMT, structured language modelling)
Alex Clark (Language theoretic models of grammar induction)
Trevor Cohn (Machine learning, SMT, grammar induction)
Yang Liu (SMT)

**Post-doc**

Adam Lopez (Large scale SMT, formal models of MT)

**PhD Students**

Chris Dyer (Large scale SMT)
Zhifei Li (Joshua, SMT, machine learning)

# Summary

- Scientific Merit:
  - A systematic comparison of existing syntactive approaches to SMT.
  - An empirical study of how constituency if useful in SMT.
  - An evaluation of existing theories of grammar induction in a practical application (end-to-end evaluation).
- Potential Impact:
  - Better MT systems, for more languages, across a range of domains.
  - More accessible high performance translation models for researchers all over the world.
- Feasibility:
  - A great team with a wide range of both theoretical and practical experience
  - Incremental plan without any deal breaking dependencies.
- Novelty:
  - First attempt at large scale unsupervised synchronous grammar induction.
  - First study seeking to compare and understand the impact of synchronous structure on translation performance.