Models of Synchronous Grammar Induction for SMT

Phil Blunsom[1] **Joy Ying Zhang**[2]
Alex Clark[3] Trevor Cohn[4] Chris Dyer[5]
Zhifei Li[6] Yang Liu[7] Adam Lopez[8]

[1] University of Oxford
[2] CMU
[3] Royal Holloway University
[4] University of Sheffield
[5] University of Maryland
[6] Johns Hopkins University
[7] Chinese Academy of Sciences
[8] University of Edinburgh

December 5, 2009

Workshop overview

Input:

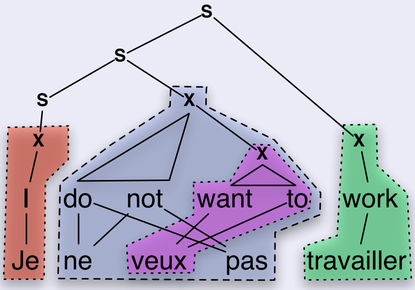
- Existing procedures for synchronous grammar extraction

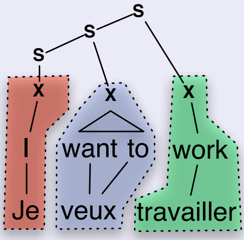
Output:

- 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

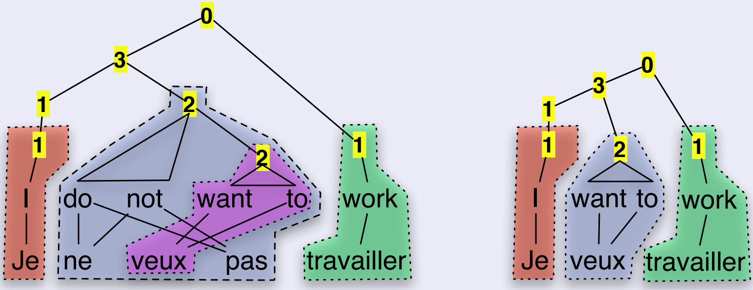
Hierarchical





Models of translation

Hierarchical



- AIM: Implement a large scale open-source synchronous constituent labelling system.
- AIM: Investigate and understand the relationship between synchronous constituency and SMT performance.

Statistical machine translation: limitations

Structural divergence between languages:

English	**The plane is faster than the train.**
Arabic	الطائرة أسرع من القطار (the-plane) (faster) (than) (the train)
Chinese	飞机 比 火车 快 (plane) (compared-to) (train) (fast)

English	**Who wrote this letter?**
Arabic	من الذي كتب هذه الرسالة؟ (function-word) (who) (wrote) (this) (the-letter)
Chinese	这封 信 是 谁 写 的 ？ (this) (letter) (be) (who) (write) (come-from) (function-word)

Inducing a STSG given an observed tree:

Hiero: An existing unsupervised extraction system

Unsupervised grammar induction

There has been significant research into monolingual grammar induction:
Constituent context is a prime indicator of constituency.

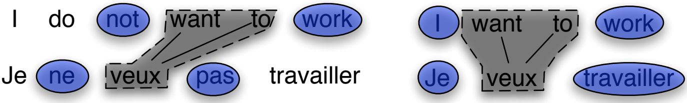
- Alexander Clark. Unsupervised induction of stochastic context-free grammars using distributional clustering, 2001
- Dan Klein and Chris Manning. A Generative Constituent-Context Model for Improved Grammar Induction, 2002

We can formalise this notion in algebraic structures

- Alexander Clark. A learnable representation for syntax using residuated lattices, 2009

Deep connections to unsupervised word sense disambiguation, thesaurus extraction etc.

Constituency and context



- Apply large scale scale clustering and topic modelling algorithms,
- 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

Phil Blunsom
Joy Ying Zhang
Alex Clark
Trevor Cohn
Chris Dyer
Zhifei Li
Yang Liu
Adam Lopez

A unique opportunity to bring together researchers operating at the coal face of SMT development with leading theoreticians in the field of formal grammar induction.

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.