Unsupervised Models of Synchronous Grammar Induction for SMT

Phil Blunsom[1]
Alex Clark[2] Trevor Cohn[3]
Chris Dyer[4] Adam Lopez[5]

[1]University of Oxford
[2]Royal Holloway University
[3]University of Sheffield
[4]University of Maryland
[5]University of Edinburgh

December 4, 2009

Statistical machine translation

Arabic → English

بغداد 1-1 (افب) - ذكرت وكالة الانباء العراقية الرسمية ان نائب رئيس
مجلس قيادة الثورة في العراق عزة ابراهيم استقبل اليوم الاربعاء في بغداد
رئيس مجلس ادارة المركز السعودي ل- تطوير الصادرات عبد الرحمن الزامل .

↓

?

- Statistical machine translation: Learn how to translate from parallel corpora.

Statistical machine translation: successes

Arabic → English

بغداد 1–1 (افب) – ذكرت وكالة الانباء العراقية الرسمية ان نائب رئيس
مجلس قيادة الثورة في العراق عزة ابراهيم استقبل اليوم الاربعاء في بغداد
رئيس مجلس ادارة المركز السعودي ل– تطوير الصادرات عبد الرحمن الزامل .

↓

Baghdad 1-1 (AFP) - official Iraqi news agency reported that vice-chairman of the revolution command council Izzat Ibrahim received in Iraq on Wednesday in Baghdad, board chairman of the Saudi center for developing exports Abdel Rahman Al-Zamil.

- Statistical machine translation: Learn how to translate from parallel corpora

Statistical machine translation: limitations

Chinese → English

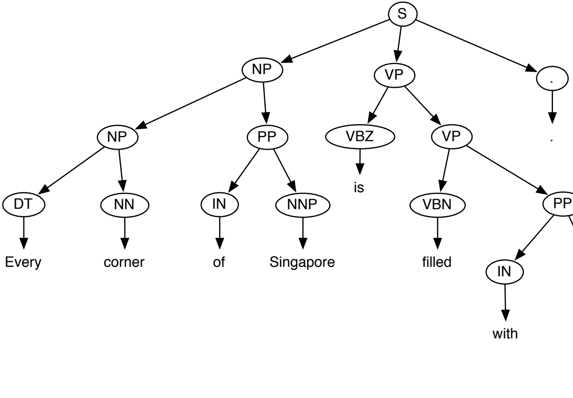
加拿大与欧盟和澳洲一样 都在十一月二十八日关闭它们的大使馆,并在本周稍早重新开放。

↓

Canada and the EU and Australia have closed on 28 November at the same as the Chinese embassy in their earlier this week, and re-opening up.

- But sometimes this doesn't work so well ...

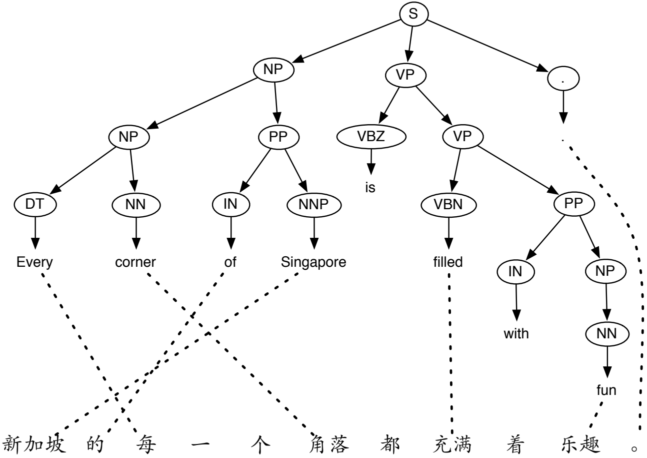
Inducing a STSG given an observed tree:



新加坡 的 每 一 个 角落 都 充满 着 乐趣 。

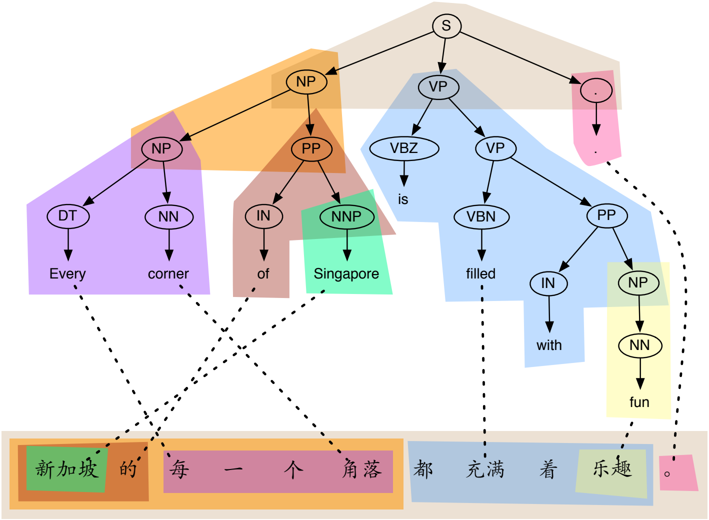
Training instance

Existing approach (Galley et al. 2004):



Step 1: word alignment

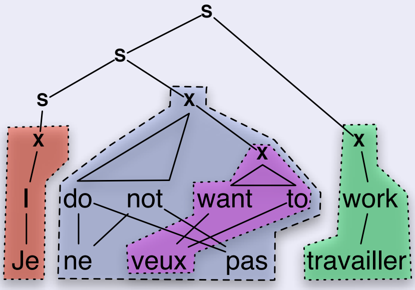
Existing approach (Galley et al. 2004):

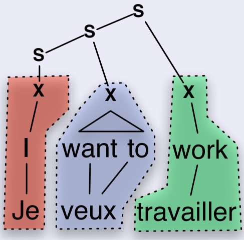


Step 2: rule extraction heuristic

Models of translation

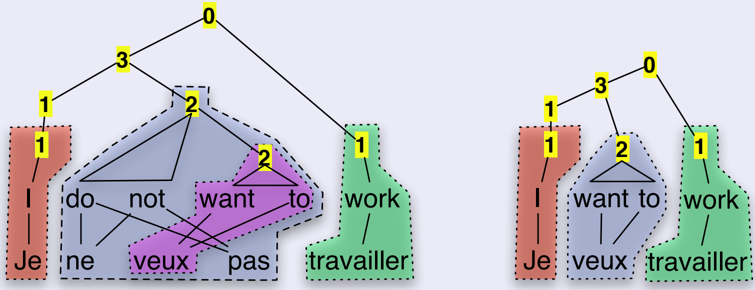
Hierarchical





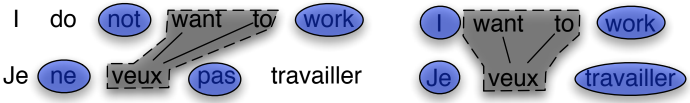
Models of translation

Hierarchical



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

Constituency and context

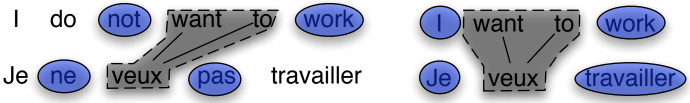


There has been significant research into monolingual grammar induction:

- 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

Constituent context is a prime indicator of constituency.

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.

The team

Phil Blunsom
Alex Clark
Trevor Cohn
Chris Dyer
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.