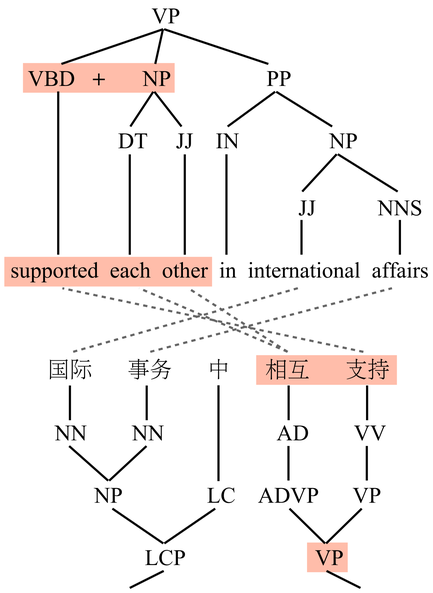
Improving Syntax-Augmented MT by Coarsening the Label Set

Carnegie Mellon University

Greg Hanneman and Alon Lavie
{ghannema, alavie}@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University

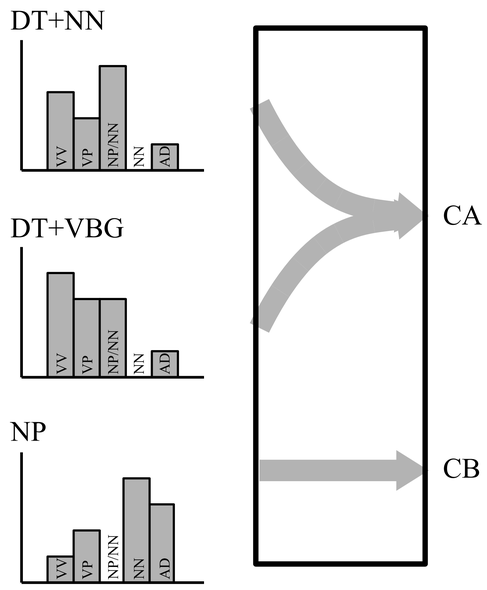
Induce Parse Trees



Extract Bilingual SAMT Rules

AD::DT+NN → [紧急]::[an emergency]
DEC+NN::DT+NN → [的 NN[1]]::[the NN[1]]
NN::DT+NN → [世界]::[the world]
PU+VV::DT+NN → [, VV[1]]::[that NN[1]]
⋮
AD::DT+VBG → [越来越]::[an increasing]
DEC+NN::DT+VBG → [的 NN[1]]::[the VBG[1]]
NN::DT+VBG → [报告]::[the reporting]
NP/NN::DT+VBG → [报刊 警告]::[this warning]
⋮
AD::NP → [不]::[no one]
AD+VV::NP → [早日 VV[1]]::[an early NN[1]]
CD::NP → [一半]::[half of them]
NP/NN::NP → [人力 资源 能力]::[human capacity]

Collapse Labels



Reduce to Monolingual SAMT

AD::**CA** → **[紧急]::[an emergency]**
DEC+NN::**CA** → **[的 CX[1]]::[the CX[1]]**
NN::**CA** → **[世界]::[the world]**
PU+VV::**CA** → **[, CX[1]]::[that CX[1]]**
⋮
AD::**CA** → **[越来越]::[an increasing]**
DEC+NN::**CA** → **[的 CY[1]]::[the CY[1]]**
NN::**CA** → **[报告]::[the reporting]**
NP/NN::**CA** → **[报刊 警告]::[this warning]**
⋮
AD::**CB** → **[不]::[no one]**
AD+VV::**CB** → **[早日 CX[1]]::[an early CX[1]]**
CD::**CB** → **[一半]::[half of them]**
NP/NN::**CB** → **[人力 资源 能力]::[human capacity]**

Label Collapsing

- Represent each target label as a distribution over source labels
- Iteratively combine pair of labels with the most similar distributions (L_1)

$$d(t_1, t_2) = \sum_{s \in S} \left| P(s|t_1) - P(s|t_2) \right|$$

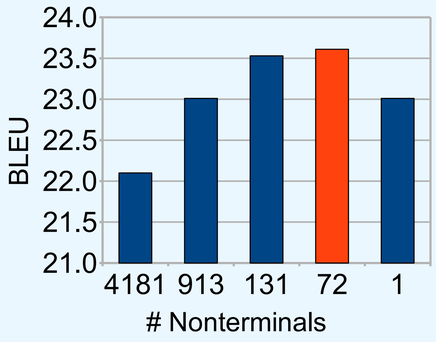
- We examine three stopping points, plus SAMT and Hiero baselines

Experiments

- FBIS Chinese–English corpus
- Tune on NIST MT 2006; test on NIST MT 2003 and 2008

Results

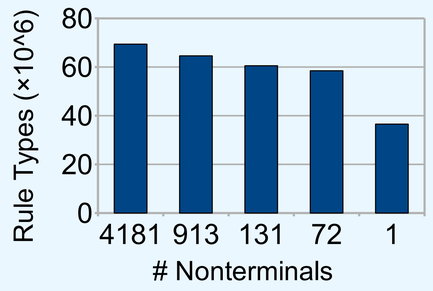
- +1.1 to +1.5 BLEU over SAMT
- +0.0 to +0.6 BLEU over Hiero
- Significantly better than both on MT 2008 according to BLEU, METEOR, and TER



Analysis

Reduced Ambiguity

- 15.7% fewer rule types extracted, but same reordering patterns covered
- Rule pattern
  X → [X[1] 的 X[2]]::[the X[2] of X[1]]
  goes from 1330 to 63 labeled variants



Reduced Sparsity

- Fraction of label types seen <100 times drops from 47% to 26%
- More syntactic rules used in output than in both SAMT and Hiero