Soft Syntactic Constraints for Hierarchical Phrase-based Translation Using Latent Syntactic Distributions

By Z. Huang, M. Čmejrek, and B. Zhou
Presented by Austin Matthews

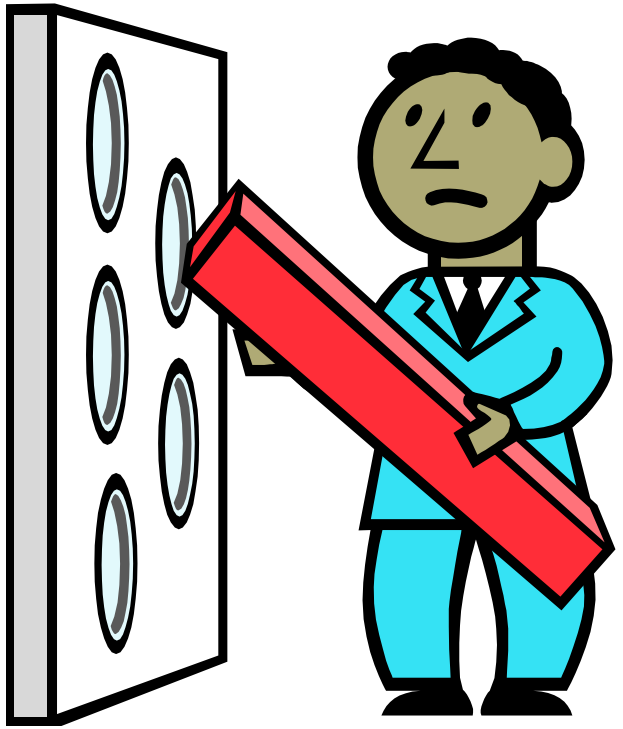
4/25/2012

The Goal

- Integrate syntax into MT without being overly constraining, but without losing linguistic guidance.
- Develop a method to determine syntactic similarity between tag sequences.

Prior Work

- Hiero, no syntax
  - Try to fit any phrase into any hole
- Zollman and Venugopal
  - Annotate phrase pairs with tree bank categories
  - Require exact match to substitute



Prior Work

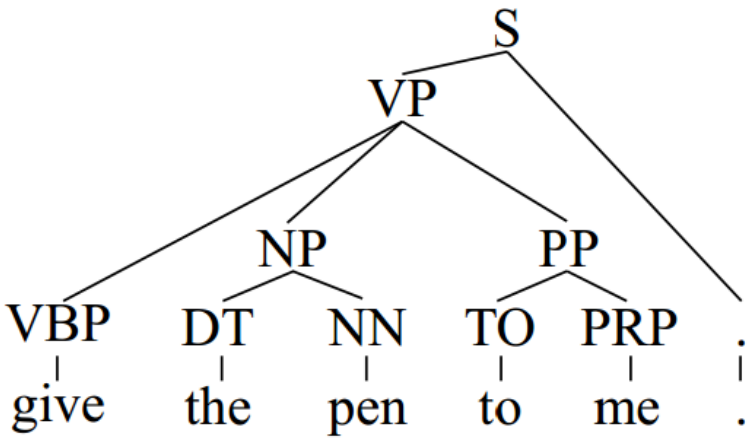
- 我爬上X_1^{NP}。→ I climb X_1^{NP}.
- 长城 → the Great Wall (NP)
- 楼梯 → the stairs (DT NN)

- Can't plug in "the stairs"!

The Idea

- Extract *tag sequences* from the data
- Each X non-terminal can be annotated with a *distribution* over the tag sequences
- I am reading …

Tag Sequence	Probability
NP	0.40
DT NN	0.35
DT NN NN	0.25



- Here ‘give the pen’ is dominated exactly and minimally by VBP NP

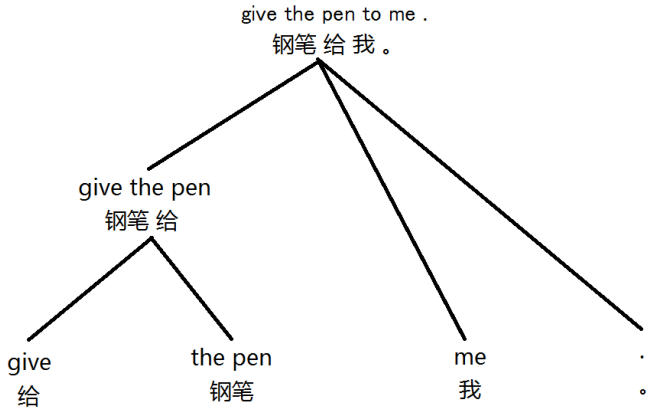
The Idea

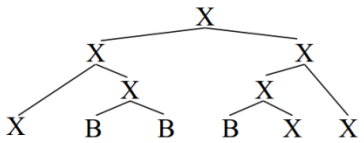
- Represent tag sequences with latent syntactic categories
- For each sequence, compute distribution over latent categories
- Can now compute similarity via dot product!
- Feature: negative log of dot product

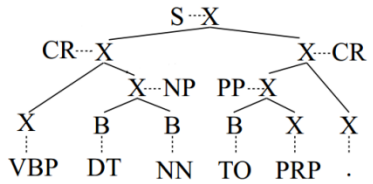
- Example:
- "NP VP" = {0.5, 0.2, 0.3} might mean NP VP acts as:
  - a VP 50% of the time
    - He gives me the pen.
  - a JP 20% of the time
    - It's a time telling device.
  - A NP 30% of the time
    - Him running bothers me.

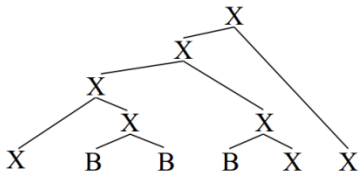
Rule Extraction

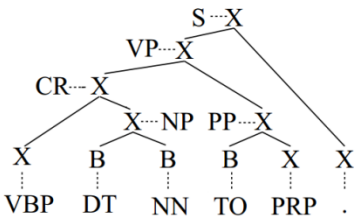
- Lemma: Two maximal phrase pairs are either disjoint, or one is fully contained within the other
- Exploit this to build a tree of maximal phrase pairs
- Binarize into a forest
- Annotate nodes with constituent labels







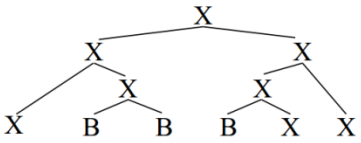


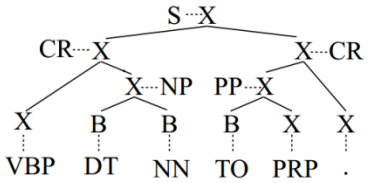


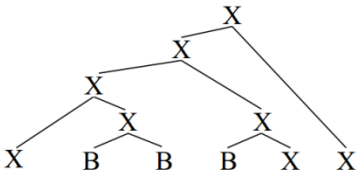
X: Phrase, B: Non-phrase

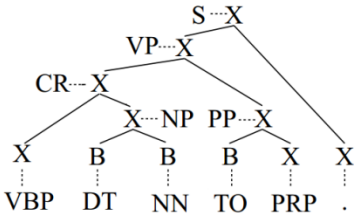
Inducing Latent Syntax

- For each X and B non-terminal, split them into {2,4,8,16} non-terminals and learn their category distribution.
- Parameter tuning done through inside-outside









Experiments

- EN→DE
  - Europarl (~300k TUs)
  - Average 15 tokens/TU
- EN→ZH
  - Travel domain (~500k TUs)
  - Average 6 tokens/TU
- Hiero baseline with 4-gram LM

- You will be aware from the press and television that there have been a number of bomb explosions and killings in Sri Lanka.
- Wie Sie sicher aus der Presse und dem Fernsehen wissen, gab es in Sri Lanka mehrere Bombenexplosionen mit zahlreichen Toten.

- Give the pen to me .
- 钢笔 给 我 。

Experiments

- Found 8.3M rules for EN→DE and 9.7M for EN→ZH
- ~180k unique tag sequences

EN→DE	Baseline	Syntax	Δ
Train	16.26	17.06	0.80
Test	16.41	17.01	0.60

EN→ZH	Baseline	Syntax	Δ
Train	46.47	47.39	0.92
Test	45.45	45.86	0.41

Evaluating Similarity Metric

	Very similar	Not so similar	Very dissimilar
DT JJ NN	DT NN DT JJ JJ NN DT ADJP NN	DT JJ JJ NML NN DT JJ CC INTJ VB DT NN NN JJ	PP NP NN NN CD VP RB NP IN CD
VP	VB VB RB VB PP VB DT DT NN	VP PP JJ NN VB NN NN VB VB RB IN JJ	JJ NN TO VP JJ WHNP DT NN IN INTJ NP
ADJP	JJ PDT JJ RB JJ	ADJP JJ JJ CC ADJP VB JJ JJ ADVP WHNP JJ	ADJP IN NP JJ AUX RB ADJP ADJP VP

Discussion

- Low margin of improvement
- Could show a phrase-based baseline
- Better to collapse similar categories top-down, or to build categories bottom-up?