Statistical NLP
Spring 2008

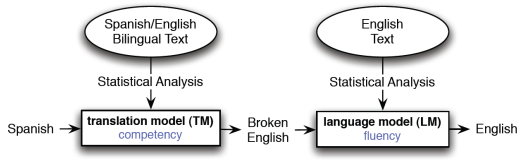
University of California
Berkeley

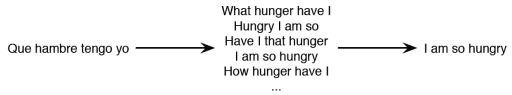
Lecture 25: Syntactic Translation

Dan Klein – UC Berkeley

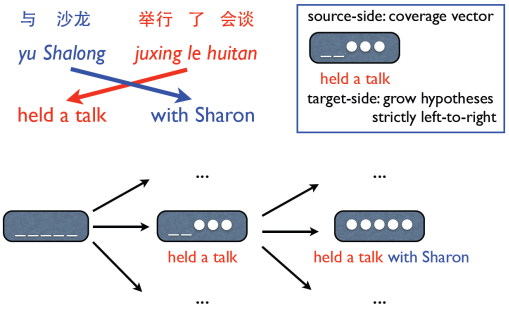
Slides from Liang Huang and Jonathan May

MT Overview





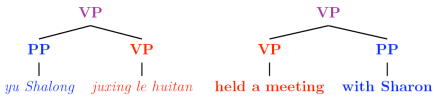
Phrase-Based MT



Syntax-Based MT

- synchronous context-free grammars (SCFGs)
  - context-free grammar in two dimensions
  - generating pairs of strings/trees simultaneously
  - co-indexed nonterminal further rewritten as a unit

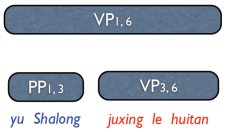
$$\begin{array}{lcll} \text{VP} & \rightarrow & \text{PP}^{(1)}\ \text{VP}^{(2)}, & \text{VP}^{(2)}\ \text{PP}^{(1)} \\ \text{VP} & \rightarrow & \textit{juxing le huitan}, & \text{held a meeting} \\ \text{PP} & \rightarrow & \textit{yu Shalong}, & \text{with Sharon} \end{array}$$



Translation by Parsing

- translation with SCFGs => monolingual parsing
- parse the source input with the source projection
  - build the corresponding target sub-strings in parallel

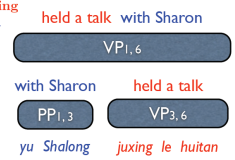
$$\begin{array}{lcl} \text{VP} & \rightarrow & \text{PP}^{(1)}\ \text{VP}^{(2)}, \\ \text{VP} & \rightarrow & \textit{juxing le huitan}, \\ \text{PP} & \rightarrow & \textit{yu Shalong}, \end{array}$$



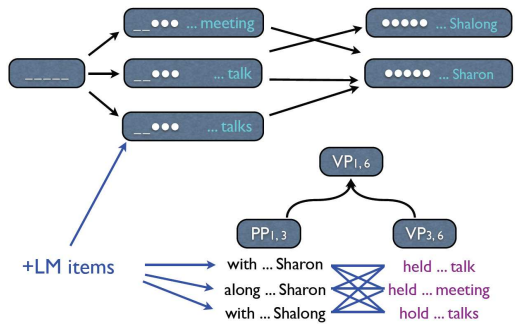
Translation by Parsing

- translation with SCFGs => monolingual parsing
- parse the source input with the source projection
  - build the corresponding target sub-strings in parallel

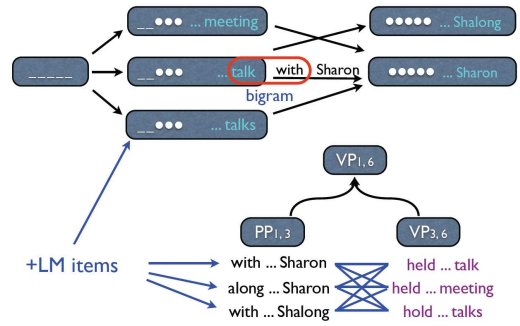
$$\begin{array}{lcll} \text{VP} & \rightarrow & \text{PP}^{(1)}\ \text{VP}^{(2)}, & \text{VP}^{(2)}\ \text{PP}^{(1)} \\ \text{VP} & \rightarrow & \textit{juxing le huitan}, & \text{held a meeting} \\ \text{PP} & \rightarrow & \textit{yu Shalong}, & \text{with Sharon} \end{array}$$



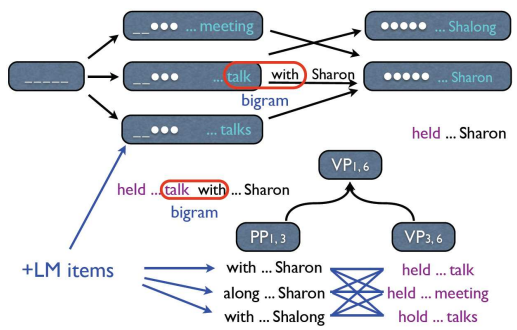
Compact Forests



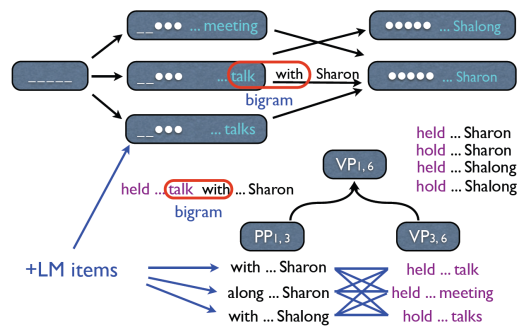
Compact Forests



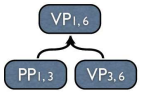
Compact Forests



Compact Forests



Language Model Costs



non-monotonic grid due to LM combo costs

	$(\mathrm{PP}^{\mathrm{with}\star\mathrm{Sharon}}_{1,3})$	$(\mathrm{PP}^{\mathrm{along}\star\mathrm{Sharon}}_{1,3})$	$(\mathrm{PP}^{\mathrm{with}\star\mathrm{Shalong}}_{1,3})$
	1.0	3.0	8.0
$(\mathrm{VP}^{\mathrm{held}\star\mathrm{meeting}}_{3,6})$ 1.0	2.0 + 0.5	4.0 + 5.0	9.0 + 0.5
$(\mathrm{VP}^{\mathrm{held}\star\mathrm{talk}}_{3,6})$ 1.1	2.1 + 0.3	4.1 + 5.4	9.1 + 0.3
$(\mathrm{VP}^{\mathrm{hold}\star\mathrm{conference}}_{3,6})$ 3.5	4.5 + 0.6	6.5 + 10.5	11.5 + 0.6

Language Model Costs

k-best parsing
(Huang and Chiang, 2005)

- a priority queue of candidates
- extract the best candidate
- push the two successors

	$(\mathrm{PP}^{\mathrm{with}\star\mathrm{Sharon}}_{1,3})$	$(\mathrm{PP}^{\mathrm{along}\star\mathrm{Sharon}}_{1,3})$	$(\mathrm{PP}^{\mathrm{with}\star\mathrm{Shalong}}_{1,3})$
	1.0	3.0	8.0
$(\mathrm{VP}^{\mathrm{held}\star\mathrm{meeting}}_{3,6})$ 1.0	2.5	9.0	9.5
$(\mathrm{VP}^{\mathrm{held}\star\mathrm{talk}}_{3,6})$ 1.1	2.4	9.5	9.4
$(\mathrm{VP}^{\mathrm{hold}\star\mathrm{conference}}_{3,6})$ 3.5	5.1	17.0	12.1

Language Model Costs

[Huang and Chiang 06]

items are popped out-of-order

solution: keep a buffer of pop-ups

2.5 2.4 5.1

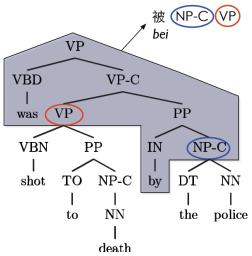
finally re-sort the buffer and return inorder:

2.4 2.5 5.1

	$(\mathrm{PP}_{1,3}^{\text{with} \star \text{Sharon}})$	$(\mathrm{PP}_{1,3}^{\text{along} \star \text{Sharon}})$	$(\mathrm{PP}_{1,3}^{\text{with} \star \text{Shalong}})$
	1.0	3.0	8.0
$(\mathrm{VP}_{3,6}^{\text{held} \star \text{meeting}})$	1.0	2.5	9.0	9.5
$(\mathrm{VP}_{3,6}^{\text{held} \star \text{talk}})$	1.1	2.4	9.5	9.4
$(\mathrm{VP}_{3,6}^{\text{hold} \star \text{conference}})$	3.5	5.1	17.0	12.1

Learning MT Grammars

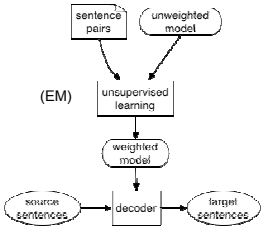
- syntax-directed, English to Chinese (Huang, Knight, Joshi, 2006)
- first parse input, and then recursively transfer



synchronous tree-substitution grammars (STSG)

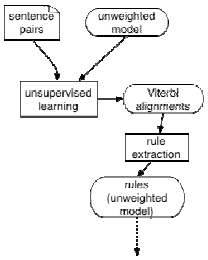
(Galley et al., 2004; Eisner, 2003)

Idealistic machine translation



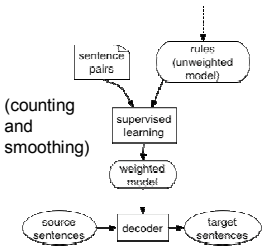
- Theorize a translation model
- Train on parallel sentences with unsupervised learning methods
- Examples: Berger et. al '94 (IBM Model 4), Wu '96 (ITG), Yamada and Knight '01

Realistic machine translation



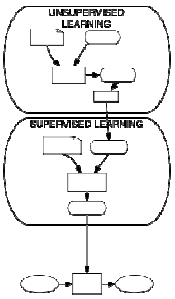
- A relatively simple model is used to learn alignments unsupervisedly
- The alignments are used to limit the exploration space of a more complicated model

Realistic machine translation



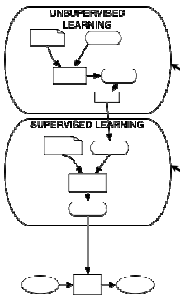
- The complicated model is trained on a smaller set of sentence pairs
- Counting and smoothing is fairly quick

Realistic machine translation



- Empirically shown to be better than idealistic approach
- Scales well
- Examples: Och & Ney '04, Chiang '05, Galley et. al '06

Realistic machine translation

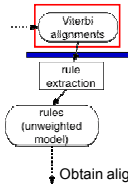


word model (IBM model 4)

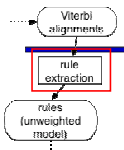
syntax model (Galley, et. al '04)

- New research is on advanced models, but older models are used as bootstrap
- Can we add the power in our supervised model to the unsupervised component?

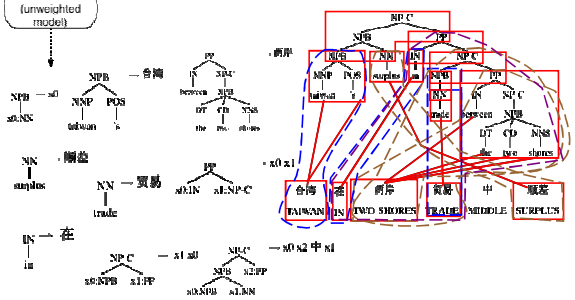
Extracting syntactic rules



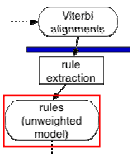
Extracting syntactic rules

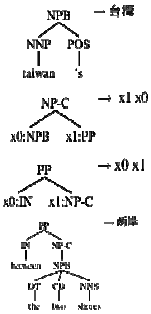


Extract rules (Galley et. al. '04, '06)



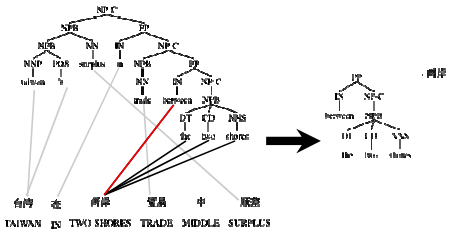
Rules can...





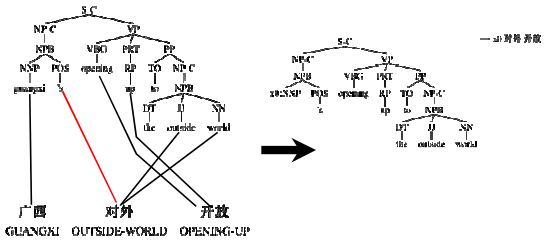
- capture phrasal translation
- reorder parts of the tree
- traverse the tree without reordering
- insert (and delete) words

Bad alignments make bad rules



This isn't very good, but let's look at a worse example...

Sometimes they're really bad

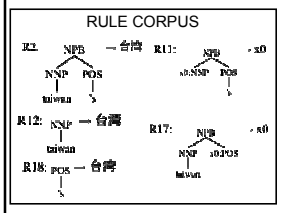


One bad link makes a totally unusable rule!

Rule Learning?

= ?

RULE CORPUS



R2 | R11 R12 | R17 R18

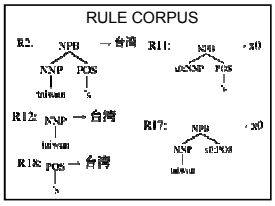
f g h

=

103 possible rules

EM size bias

RULE CORPUS



R2 | R11 R12 | R17 R18

more useful

favored

- EM favors shorter derivations: fewer chances local probability factors
- This, then, tends to favor larger rules, generally the opposite of what we want
- Various solutions: e.g. minimal rule extraction, size penalties (May and Knight, 07)

Other Important Issues

- Rule composition
- Tree transforms (binarization, etc)
- Crazy pruning

Results

From [DeNeefe et al 07]

Experiment	Chinese Dev	Chinese Test	Arabic Dev	Arabic Test
Baseline ATS	34.94	32.83	50.46	50.52
Baseline GHKM (minimal only)	38.02	37.67	49.34	49.99
GHKM composed size 2	40.24	39.75	50.76	50.94
GHKM composed size 3	40.95	40.44	51.56	51.48
GHKM composed size 4	41.36	40.69	51.60	51.71
GHKM minimal + SPMT model 1	39.78	39.16	50.17	51.27
GHKM composed + SPMT model 1	42.04	41.07	51.73	51.53
With binarization	42.17	41.26	52.50	51.79