"Learning Hierarchical Translation Structure with Linguistic Annotations"

Markos Mylonakis and Khalil Sima'an
ILLC, University of Amsterdam

Presented by Greg Hanneman

11-734: Advanced MT Seminar
March 21, 2012

Carnegie Mellon

Setup

- Hierarchical MT systems:

Constrained to linguistic syntax	vs.	**Soft syntactic constraints**
GHKM tree-to-string tree transducers		Hiero SAMT Marton and Resnik

Hierarchical Reordering SCFG

- All rules have one of the following forms:

Monotonic translation

$$A \rightarrow [B\ C]::[B\ C]$$

$$A^{L} \rightarrow [B\ C]::[B\ C]$$

$$A^{R} \rightarrow [B\ C]::[B\ C]$$

Reordered translation

$$A \rightarrow [B^{L}\ C^{R}]::[C^{R}\ B^{L}]$$

$$A^{L} \rightarrow [B^{L}\ C^{R}]::[C^{R}\ B^{L}]$$

$$A^{R} \rightarrow [B^{L}\ C^{R}]::[C^{R}\ B^{L}]$$

Phrase pair emission

$$A \rightarrow [A_{P}]::[A_{P}]$$

$$A^{L} \rightarrow [A_{P}]::[A_{P}]$$

$$A^{R} \rightarrow [A_{P}]::[A_{P}]$$

Phrase pair generation

$$A_{P} \rightarrow [x]::[y]$$

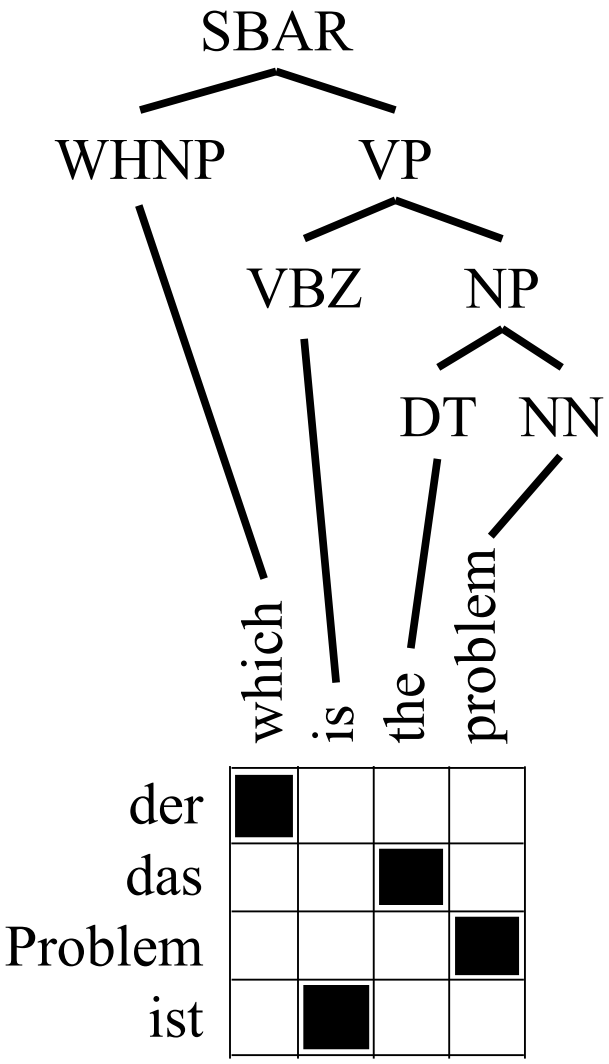
$$A_{P}^{L} \rightarrow [x]::[y]$$

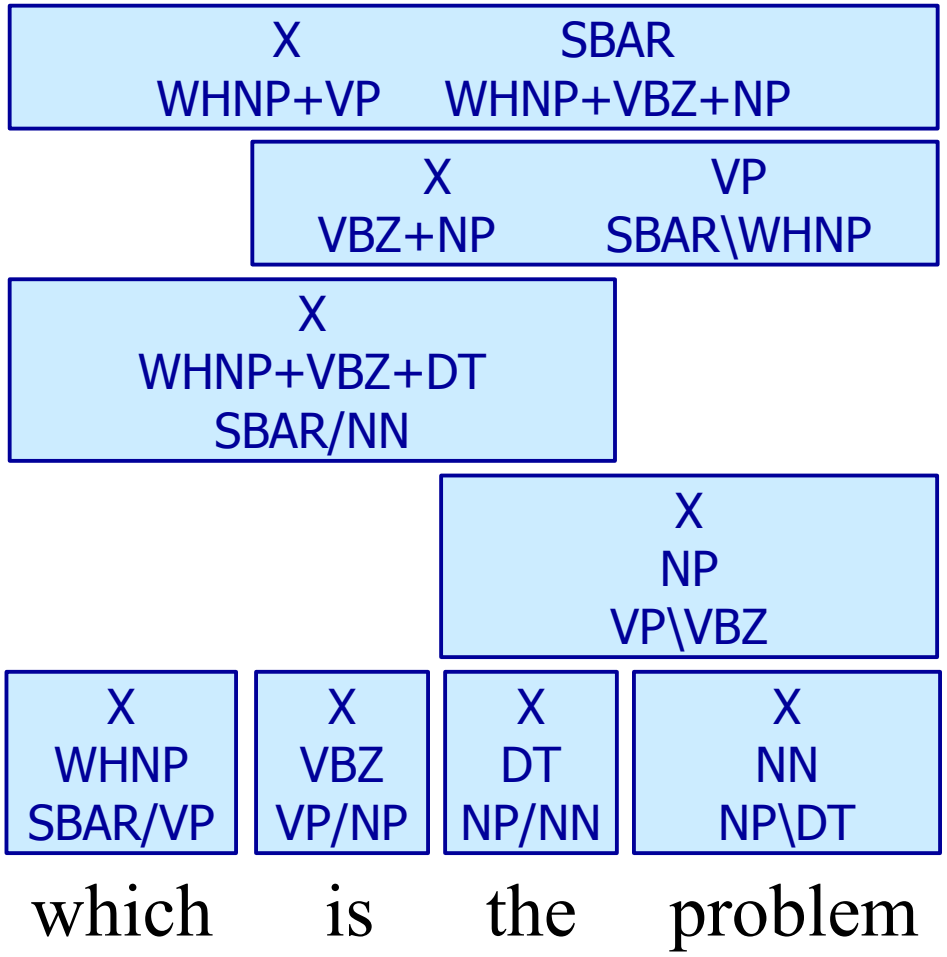
$$A_{P}^{R} \rightarrow [x]::[y]$$

Extracting Rules

- Word-align parallel corpus
- Parse source side (Charniak)
- Construct all possible labels for consistently aligned spans (SAMT-style)
- Extract (minimal?) rules

Extracting Rules





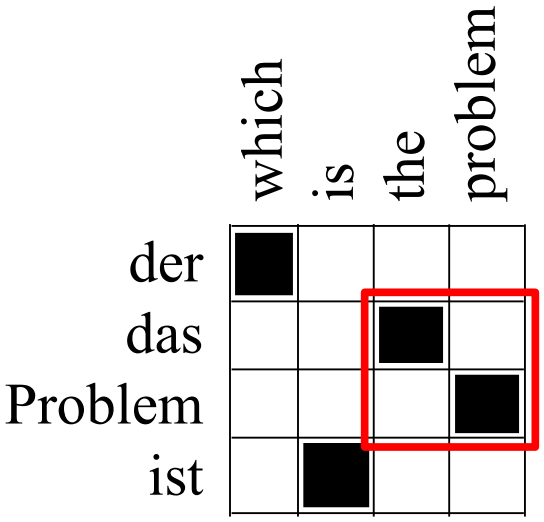
Extracting Rules

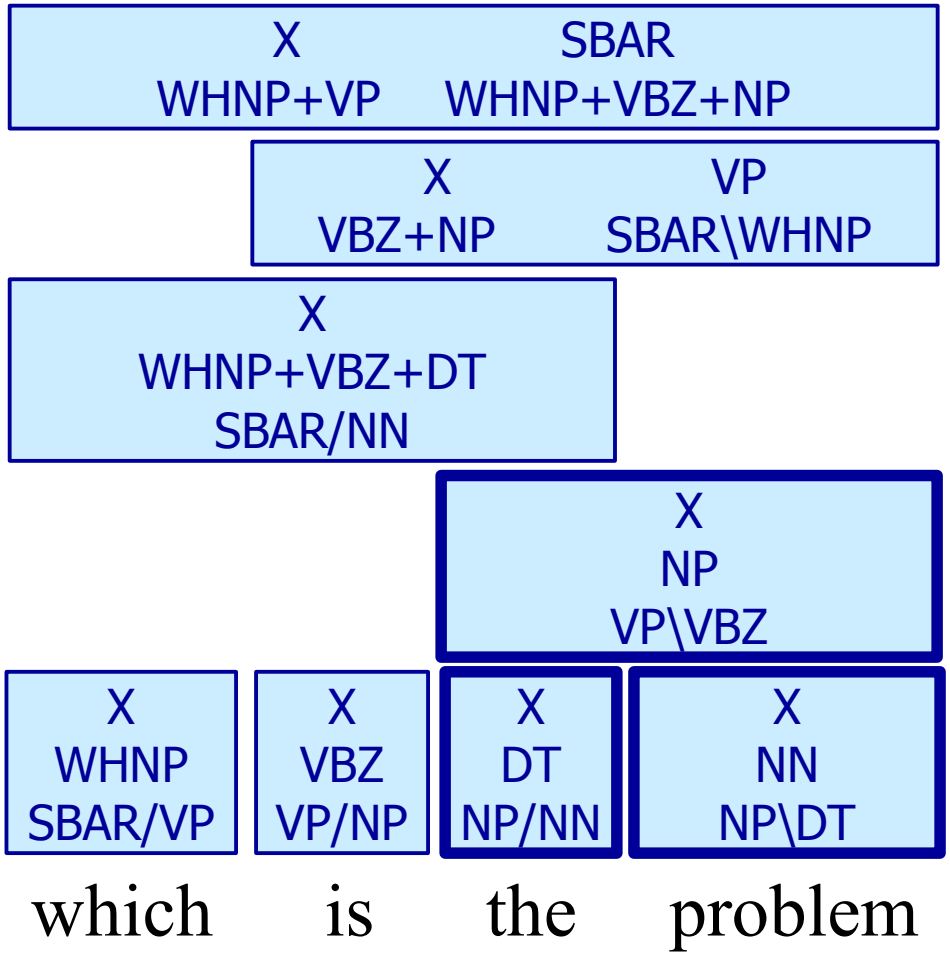
$NP^R \rightarrow [DT\ NN]::[DT\ NN]$

$VP\backslash VBZ^R \rightarrow [X\ X]::[X\ X]$

$NP^R \rightarrow [NP_P{}^R]::[NP_P{}^R]$

$NP_P{}^R \rightarrow$ [the problem]::
[das Problem]





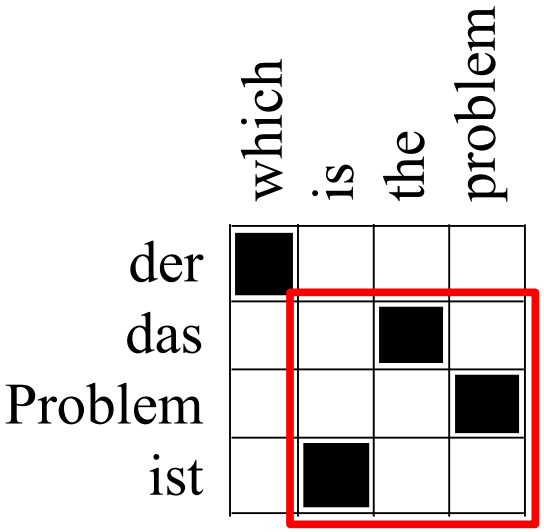
Extracting Rules

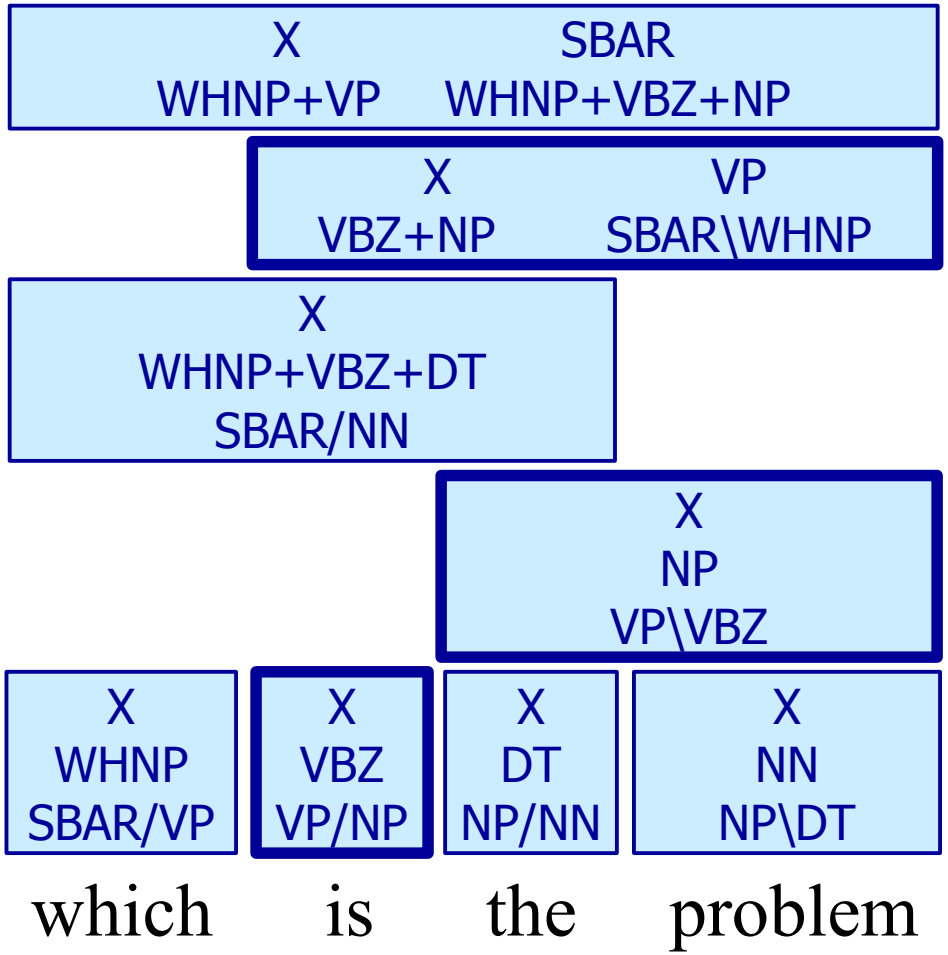
$VP \rightarrow [VBZ^L\ NP^R]::$
$[NP^R\ VBZ^L]$

$X \rightarrow [VBZ^L\ X^R]::[X^R\ VBZ^L]$

$VBZ{+}NP \rightarrow [VBZ{+}NP_P]::$
$[VBZ{+}NP_P]$

…





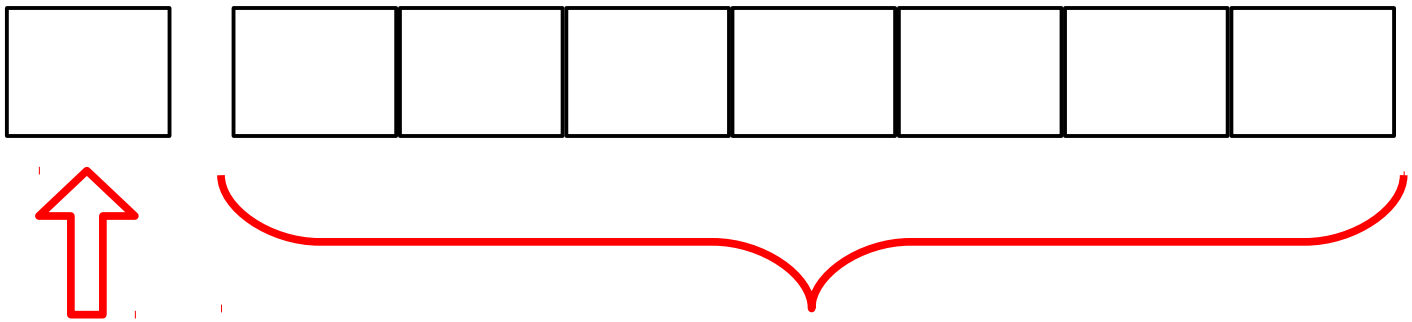
Scoring Rules

- Phrase pair rules $A_P \rightarrow [x]::[y]$
  - P(x, y | A_P) estimated from label charts
  - Standard P(x | y) and P(y | x)
  - Standard P_{lex}(x | y) and P_{lex}(y | x)
  - Word penalty f(|y|)

Scoring Rules

- Hierarchical rules
  - Estimate P(RHS | LHS), but not using MLE

Cross-Validating EM



Maximize likelihood of this portion of the training data...

... using the model that comes from the rest of the data

  - Reordering count

So That's Zillions of Rules...

- Rules appearing in only one partition of training data get ignored
- Rules below a minimum expected count in CV-EM get removed
- Decoder restricted to label chart of an input sentence
- Decoder cells have separate bins for each nonterminal

Scoring Derivations

- Probability of a derivation =
  - Language model probability ×
  - Product of scores for each phrase pair ×
  - Product of scores for each hierarchical rule

  Feature weights trained using MERT

- Probability of a joint output =
  - Sum over all derivations that produce it

  No Viterbi approximation?

Experiments

- Trained on Europarl / news data
  - Very small: 200k or 400k sentence pairs!
  - En to French, German, Dutch, and Chinese
- WMT 2007 (news) test set
- Modified KN language model
  - Very small: 1 million sentences
  - Trigram

- Compare with Joshua/Hiero baseline

Characteristic Results

- 400k training sentences, BLEU

	Joshua	LTS	
French	29.58	29.83	+0.25
German	18.86	19.49	+0.63
Dutch	22.25	22.92	+0.67
Chinese	23.24	25.16	+1.92

Discussion: Contributions

- Viable syntactic grammar that's some amalgam ITG, SAMT, or Hiero
- More explicit modeling of reordering behavior by category type / context
- Scoring that uses held-out data to go beyond "count and normalize"

- Follow-up experiment: Non-X labels help

Discussion: "How You Say It"

- Syntax-based MT has "inadequate constraints"?
  Doesn't desire to soften indicate restrictive constraints?
- SCFGs have "weak independence assumptions"?
  Doesn't passing ordering info weaken a strong assumption?
- Highest-probability rule will "always" win?
  - Conditional probabilities?
  - Reordering-based features?
  - Language model?