A Smorgasbord of Features for Statistical Machine Translation

Franz Josef Och, Daniel Gildea,
Anoop Sarkar, Kenji Yamada,
Sanjeev Khudanpur, Alex Fraser,
Shankar Kumar, David Smith, Libin
Shen, Viren Jain, Katherine Eng,

Zhen Jin, Dragomir Radev

Enormous progress in MT due to statistical methods

- Enormous progress in recent years
  - TIDES MT Evaluation: ΔBLEU=4-7% per year
  - Good research systems outperform commercial-off-the-shelf systems
    - On BLEU/NIST scoring
    - Subjectively

But still many mistakes in SMT output…

- Missing content words:
  - MT: Condemns US interference in its internal affairs.
  - Human: **Ukraine** condemns US interference in its internal affairs
- Verb phrase:
  - MT: Indonesia **that oppose the presence** of foreign troops.
  - Human: Indonesia reiterated its opposition to foreign military presence.
- Wrong dependencies
  - MT: …, particularly those who cheat the audience the players.
  - Human: …, particularly those **players who cheat the audience**.
- Missing articles:
  - MT: …, he is fully able to activate team.
  - Human: … he is fully able to activate **the** team.

What NLP tools are used by best SMT system?

STD NLP TOOLS

- Named Entity tagger
- POS tagger
- Shallow parser
- Deep parser
- WordNet
- FrameNet
- …

- **USED:**
  - N-grams
  - Bilingual phrases
  - (+rule-based translation of numbers&dates)

- Can we produce better results with POS tagger/parser/…?

"Syntax for SMT"-Workshop

- 6-week NSF Workshop at JHU
- Goal:
  Improve Chinese-English SMT quality by using 'syntactic knowledge'
- Baseline system: best system from TIDES MT evaluations
  - Alignment template MT system (ISI)

Baseline system

- Alignment template MT system
  - Training corpus: 150M words per language
  - Training: Store ALL aligned phrase pairs
  - Translation: Compose 'optimal' translation using learned phrase pairs

| Treffen wir uns | nächsten Mittwoch | um halb sieben | . |
| Let's meet | next Wednesday | at six thirty | . |

Baseline System

- Log-Linear Model

$$\Pr(e|f) = p_{\lambda_1^M}(e|f) \propto \exp\left[\sum_{m=1}^{M} \lambda_m h_m(e,f)\right]$$

  – Here: small number of informative features
  – Baseline: 11 features
- Maximum BLEU training
  – [Och03; ACL]
  – Advantage: directly optimizes quality

Approach: Incremental Refinement

1. Error analysis
2. Develop feature function 'fixing' error
3. Retrain using add'l feature function
4. Evaluate on test corpus
   - If useful: add to system
5. Goto 1

**Advantage**: Building on top of strong baseline

Approach:
Rescoring of N-Best List

- Problem: How to integrate syntactic features?
  - Parser/POS-tagger are complicated tools in itself
  - Integration into MT system very hard
- Solution: Rescoring of (precomputed) n-best lists
  - No need to integrate features in DP search
  - Arbitrary dependencies:
    - Full Chinese + English Sentence, POS sequence, parse tree
    - No left-to right-constraint
  - Simple software architecture

How large are potential improvements?

- During workshop:
  - Development corpus: 993 sentences ('01 set)
  - Test corpus: 878 sentences ('02 set)
  - 1000-best list
- First best score: BLEU=31.6%
- **Oracle Translations**
  - best possible set of translations in n-best list

How large are potential improvements?
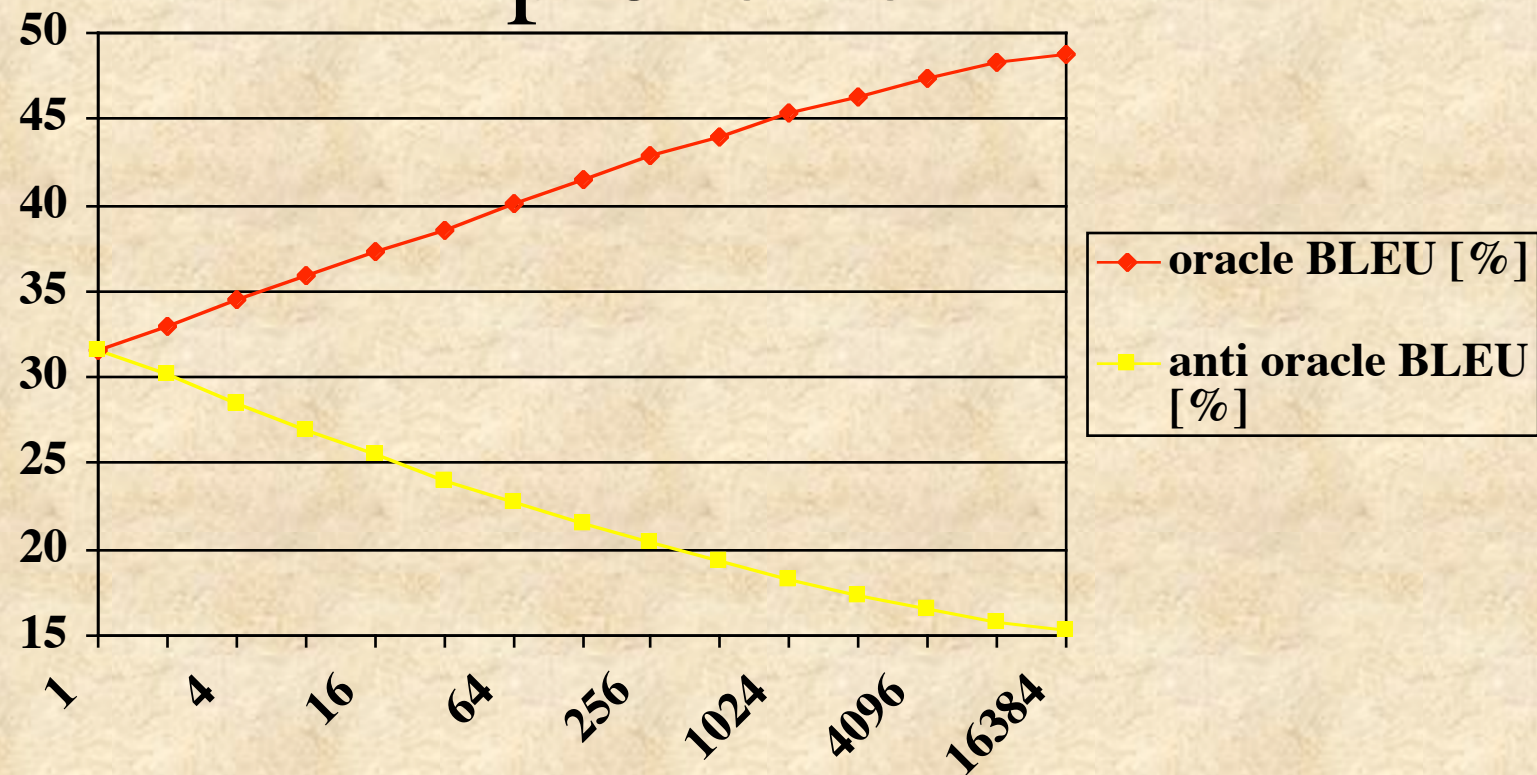
Note: 4-reference oracle too optimistic (see paper)

Syntactic Framework

- Tools
  - Chinese segmenter: LDC, Nianwen Xue
  - POS tagger: Ratnaparkhi, Nianwen Xue
  - English parser: Collins (+Charniak)
  - Chinese parser: Bikel (Upenn)
  - Chunker: fnTBL (Ngai, Florian)

- Data processed (pos-tagged/chunked/parsed)
  - Train: 1M sents (English), 70K sents (Chinese)
  - Dev/Test (n-bests): 7000 sents with 1000 bests

Feature Function Overview

- Developed 450 feature functions
  - Tree-Based
  - Tree Fragment-Based
  - Shallow: POS tags, chunker output
  - Word-Level: words and alignment

- Details: final report, project presentation slides
  `http://www.clsp.jhu.edu/ws03/groups/translate/`
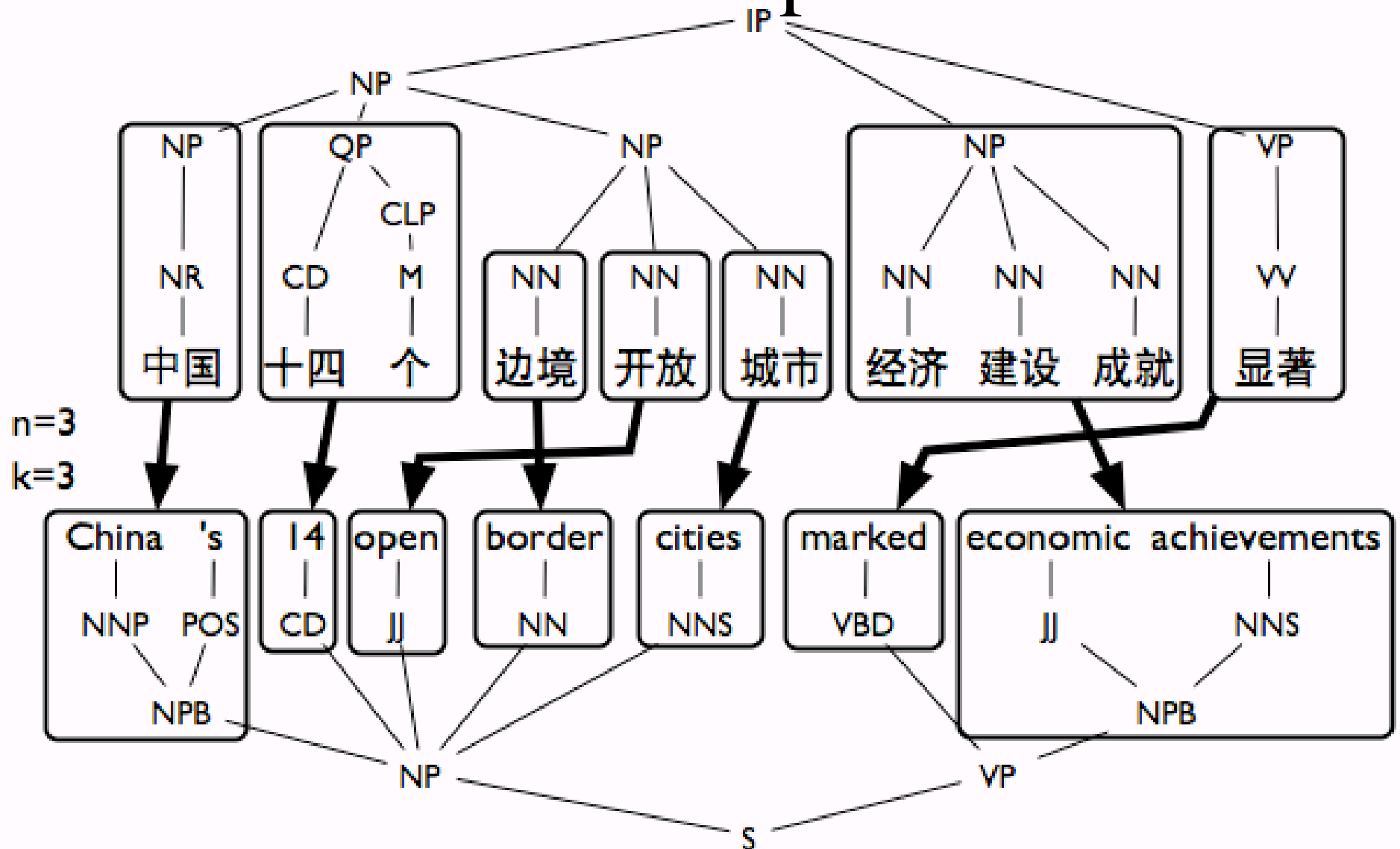
Tree-Based Features

- Tree Probability
- Tree-to-String: Project English parse tree onto Chinese string (Yamada&Knight 2001)
- Tree-to-Tree: Align trees output by both parsers node-by-node (Gildea 2003)

Result: insignificant improvement less than 0.2%

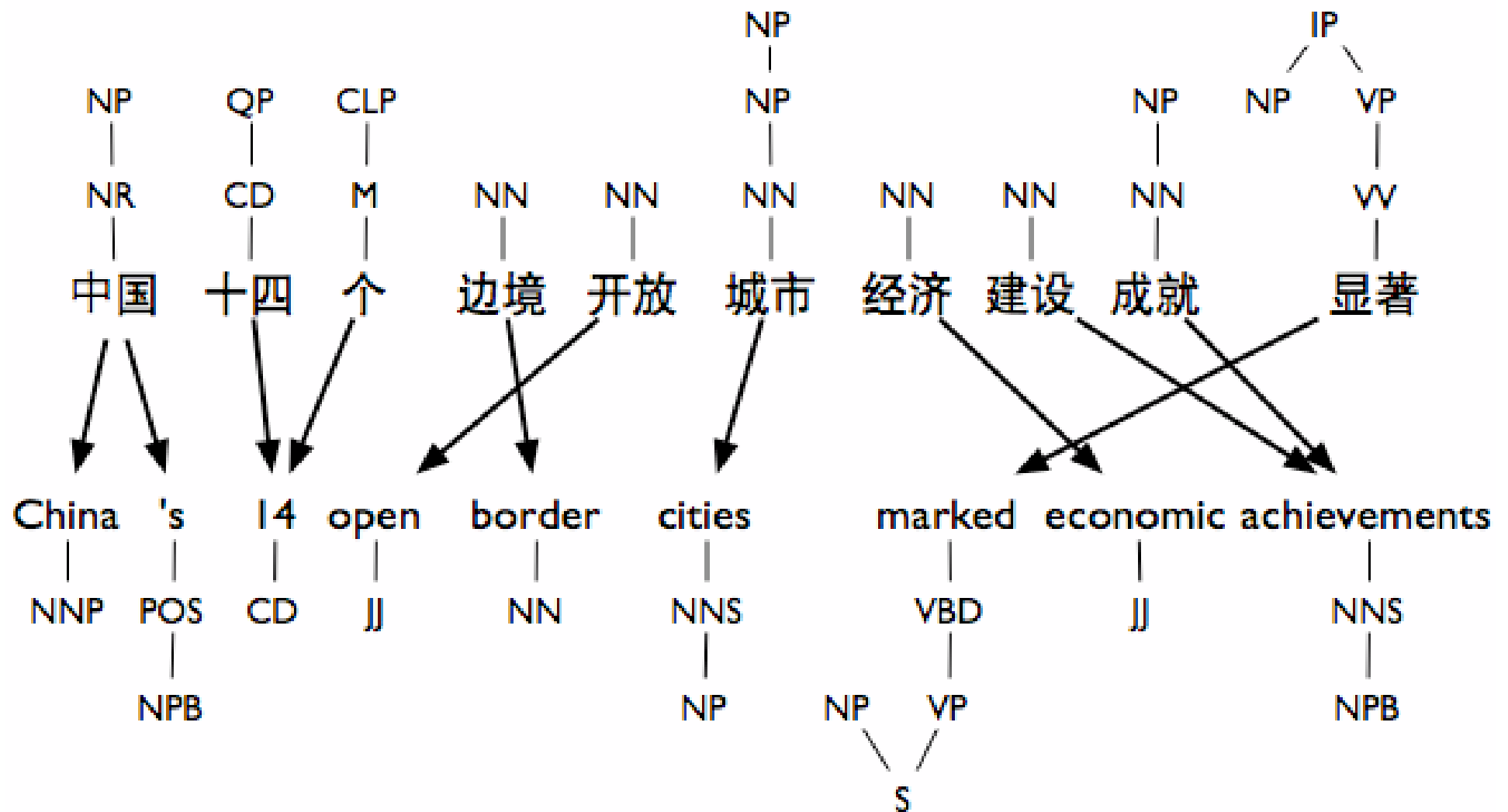**Problems**: efficiency, noisy alignments and noisy trees => **tree decomposition**

Tree Decomposition

$$h_{MarkovTreeToString} = \log(P\_TreeToString(Frag0)) + \log(P\_TreeToString(Frag1)) + \dots$$

Features From Tree Fragments

Features From Tree Fragments

- Fragment language model: unigram, bigram
- Fragment Tree-to-String Model

Result: improvement <=0.4%

Shallow Syntactic Features

Projected POS Language Model:

- Project Chinese POS to English (using alignment)
- Attach to POS symbol change in word position
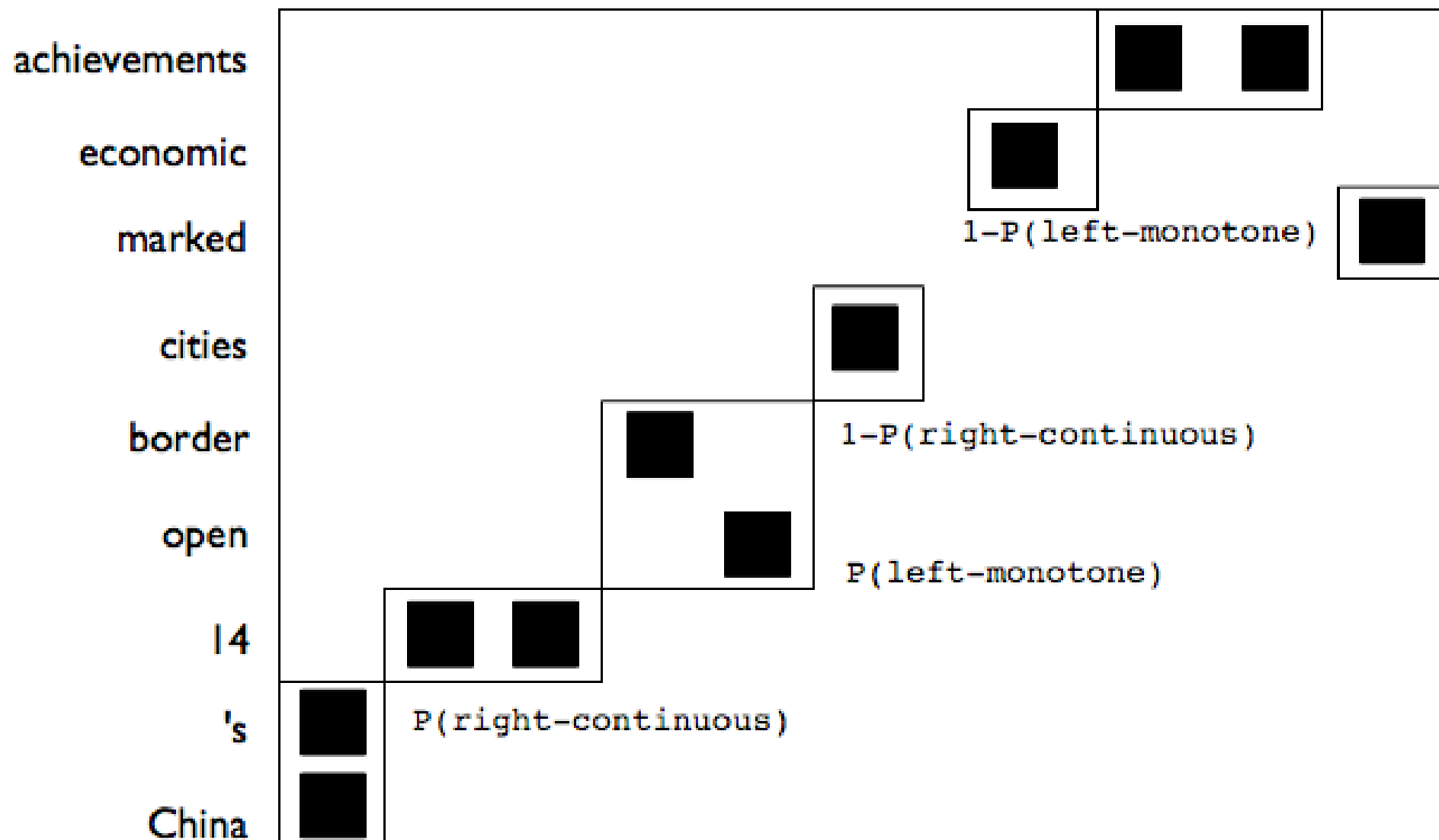- Trigram language model on resulting symbols

Example:

CD+0_M+1   NN+3   NN-1   NN+2_NN+3
 Fourteen   open   border   cities

Word/Phrase-Level

- Best features: give statistically significant improvement
- IBM Model 1 score: lexical translation probabilities w/o word order
  - P(chinese-words | english-words)
  - Sum of all alignments (no Viterbi): Triggering effect
  - Seems to fix tendency of baseline to delete content words
- Lexicalized phrase reordering model
  - Next slide

Features on Phrase Alignment

China | 's | 14 | open | border | cities | marked | economic | achievements

中国 十四 个 边境 开放 城市 经济 建设 成就 显著

P(right-continuous)

P(left-monotone)

1-P(right-continuous)

1-P(left-monotone)

Syntax for SMT - Results

- End-to-End improvement by greedy feature combination: 1.3%
  - 31.6% to 32.9%: statistically significant
  - (+ minimum Bayes risk decoding: 1.6%)

- Improvements due to:
  - Word/Phrase Level FF (>1%; statistically significant)
  - Shallow / Tree-Fragment Based (<=0.4%)
  - Tree-Based (<=0.2%)

- Conclusion: unfortunately no significant improvement using explicit syntactic analysis

Syntax - Potential Reasons for Small Improvements?

- Parsers not trained on general news text
  - ParserProb(MT output)>ParserProb(Oracle)
  - ParserProb(Oracle)>ParserProb(HumanReference)
- Parse trees often not corresponding between SL and TL
  - Many structural divergences between SL and TL
- Parsing 'bad MT output' problematic
  - Parser 'hallucinate' structures, constituents
  - In sentences without verb: noun gets analyzed as verb

Parsing/Tagging Noisy Data
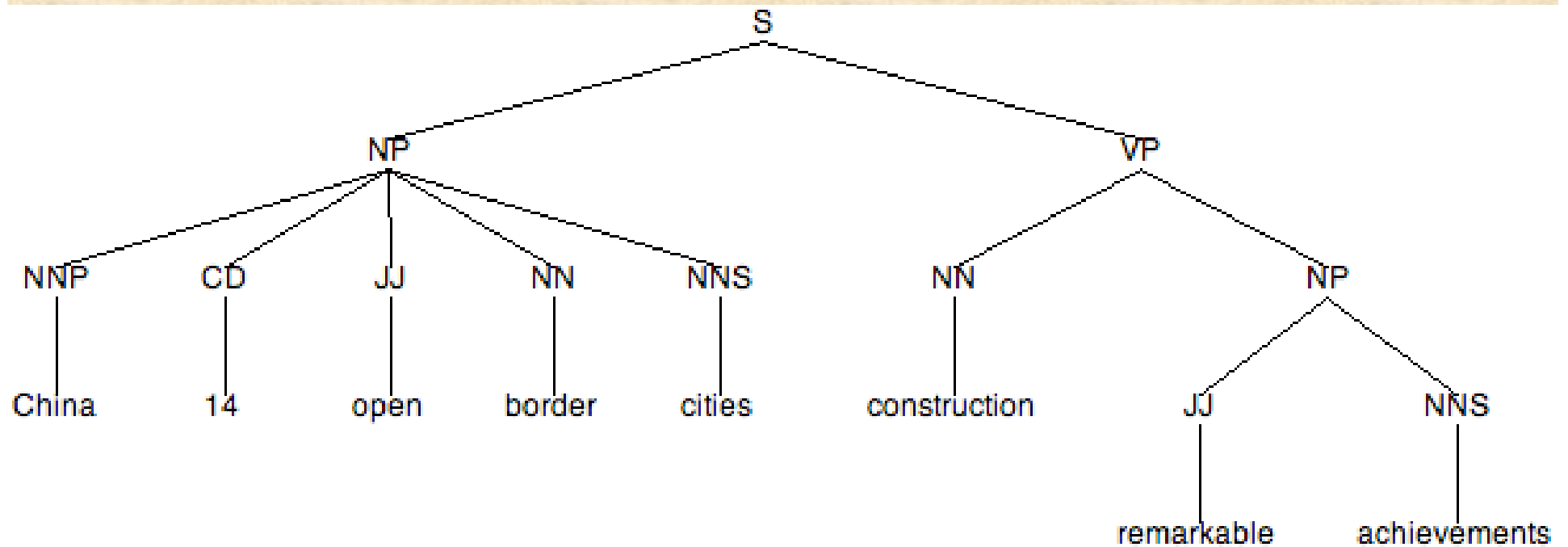
# Syntax - Potential Reasons for Small Improvements?

- Limited scalability of used framework?
  - Small Discriminative Training Corpus (993 sentences)
  - Maximum BLEU training prone to overfitting
  - Therefore: No training run on all 450 features
- Baseline system is too good?
  - Baseline MT trained on 170M words
  - Parser/Tagger trained on 1M words
- Is BLEU the right objective function for subtle improvements in syntactic quality?

Conclusions

- Discriminative reranking of N-Best lists in MT is a promising approach
  - 1.6% overall improvement on 1000-best list in 6 weeks on top of best Chinese-English MT system
- Still unclear if parsers are useful for (S)MT
  - What kind of analysis tools would be helpful?
  - B. Mercer: "**With friends like statistics, who needs linguistics?**" -- true for MT?

Round-robin (l1o-oracle) vs. optimal oracle (avBLEUr3n4)
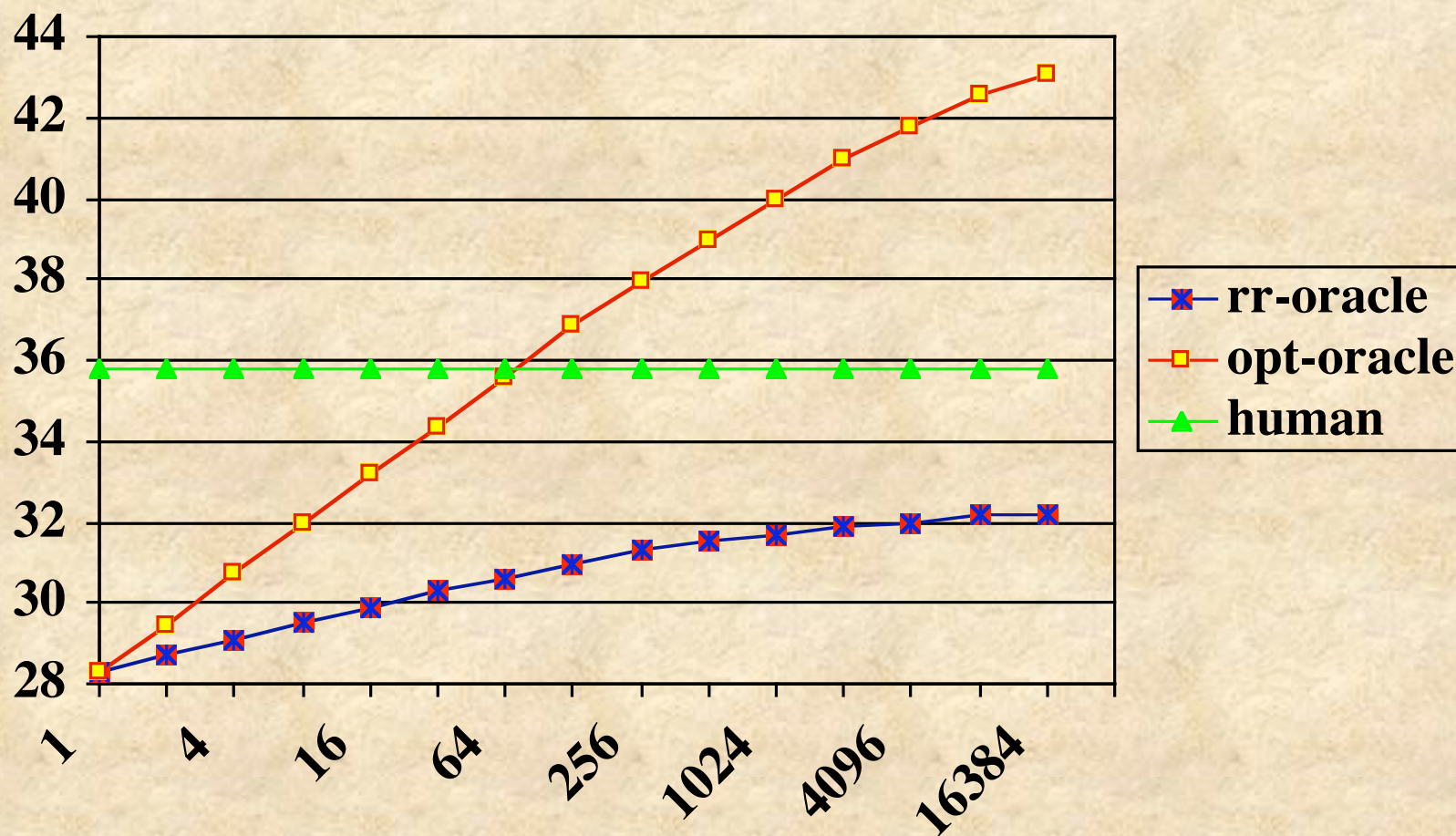
Processing Noisy Data

- Tagger tries to "fix up" ungrammatical sentences
  - China_NNP 14_CD open_JJ border_NN cities_NNS achievements_VBZ remarkable_JJ
- Same effects in parser
- **Resulting problem**: parses will look syntactically well-formed even for ill-formed sentences

Example Chinese-English

- North Korean Delegation, North Korea Has No Intention to Make Nuclear Weapons

- Seoul (Afp) - South Korean officials said that the North and South Korea ministerial-level talks between the North Korean delegation, said today that North Korea has no intention to make nuclear weapons.

- South Korean delegation spokesman Li FUNG said that North Korea, "North Korea that it was not making nuclear weapons," he said.