Automatic Category Label Coarsening for Syntax-Based Machine Translation

Greg Hanneman and **Alon Lavie**

Language Technologies Institute
Carnegie Mellon University

Fifth Workshop on Syntax and Structure in Statistical Translation
June 23, 2011

Carnegie Mellon

Motivation

- SCFG-based MT:
  - Training data annotated with constituency parse trees on both sides
  - Extract labeled SCFG rules

$$\text{A::JJ} \rightarrow [\text{bleues}]::[\text{blue}]$$

$$\text{NP::NP} \rightarrow [\text{D}^1\ \text{N}^2\ \text{A}^3]::[\text{DT}^1\ \text{JJ}^3\ \text{NNS}^2]$$

- We think syntax on both sides is best
- But joint default label set is too large

Motivation

- Labeling ambiguity:
  - Same RHS with many LHS labels

JJ::JJ → [快速]::[fast]

AD::JJ → [快速]::[fast]

JJ::RB → [快速]::[fast]

VA::JJ → [快速]::[fast]

VP::ADJP → [VV^1 VV^2]::[RB^1 VBN^2]

VP::VP → [VV^1 VV^2]::[RB^1 VBN^2]

Motivation

- Rule sparsity:
  - Label mismatch blocks rule application

VP::VP → [VV^1 了 PP^2 的 NN^3]::[VBD^1 their NN^3 PP^2]

VP::VP → [VV^1 了 PP^2 的 NN^3]::[VB^1 their NNS^3 PP^2]

✓ saw their friend from the conference

✓ see their friends from the conference

✗ **saw** their **friends** from the conference

Motivation

- Solution: modify the label set
- Preference grammars [Venugopal et al. 2009]
  - X rule specifies distribution over SAMT labels
  - Avoids score fragmentation, but original labels still used for decoding
- Soft matching constraint [Chiang 2010]
  - Substitute A::Z at B::Y with model cost subst(B, A) and subst(Y, Z)
  - Avoids application sparsity, but must tune each subst(s_1, s_2) and subst(t_1, t_2) separately

Our Approach

- Difference in translation behavior ⇒ different category labels

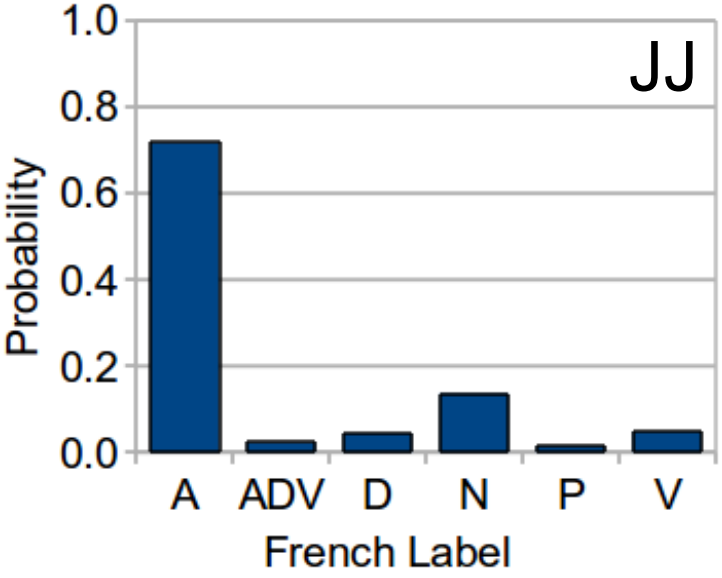
la grande voiture	the large car
la plus grande voiture	the larger car
la voiture la plus grande	the largest car

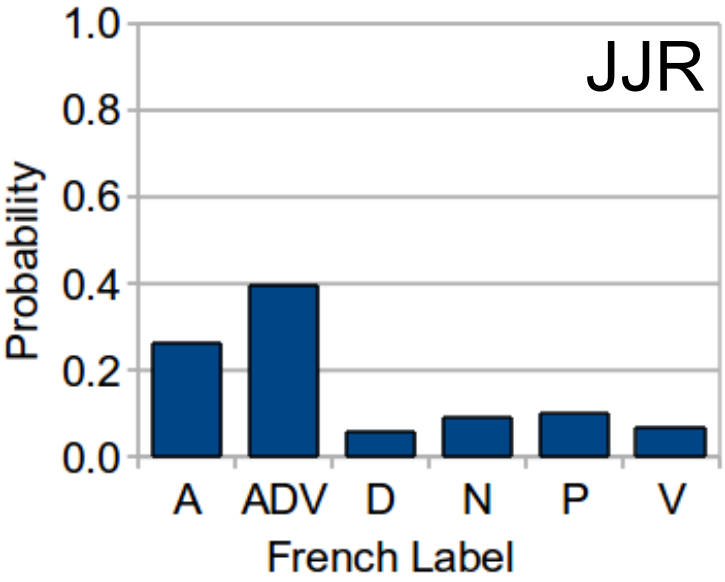
- Simple measure: how category is aligned to other language

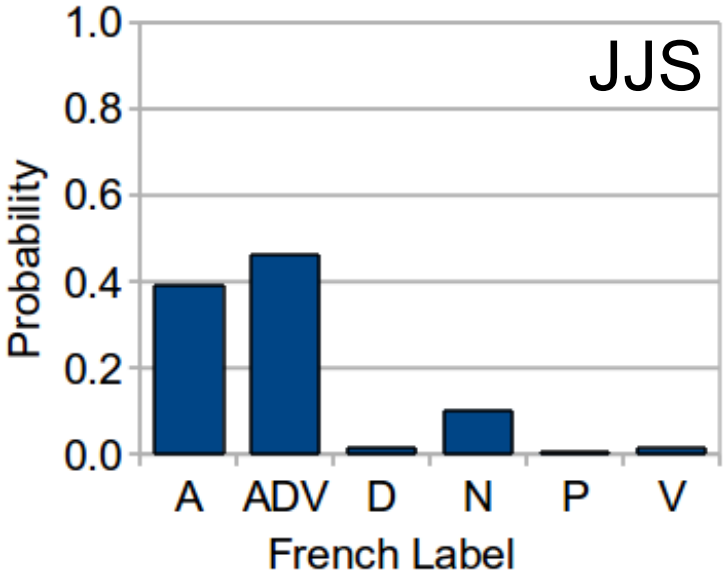
A::JJ → [grande]::[large]

AP::JJR → [plus grande]::[larger]

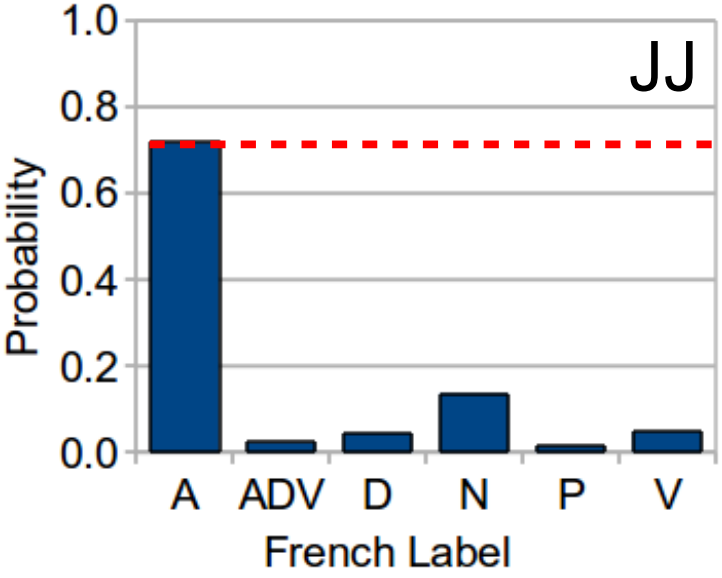
L_1 Alignment Distance

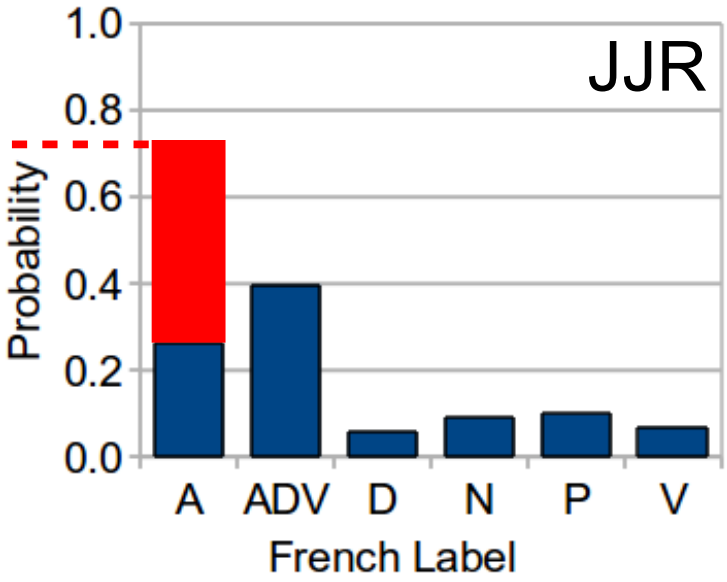


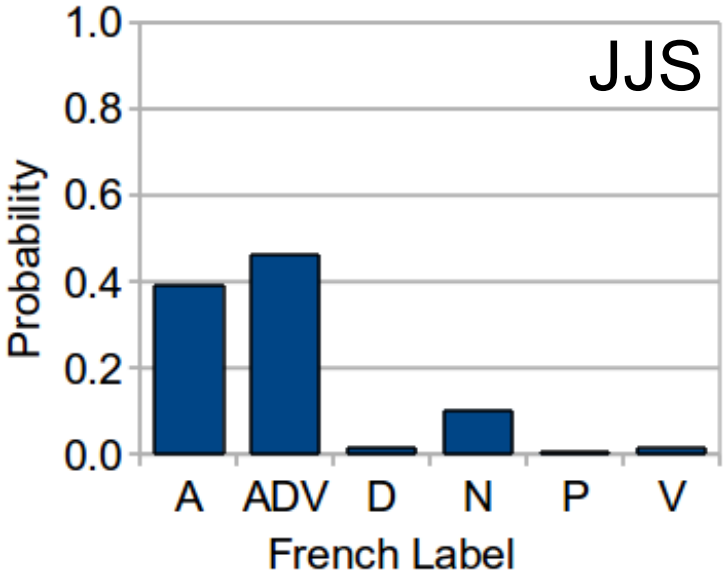




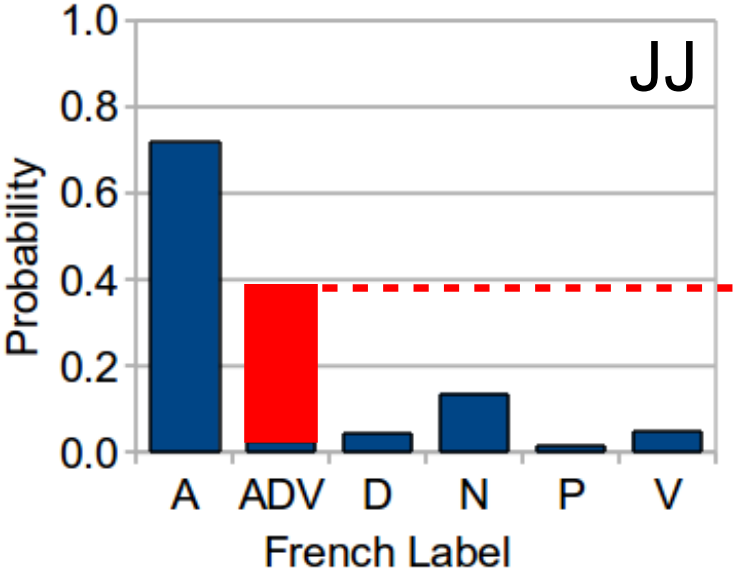
L_1 Alignment Distance

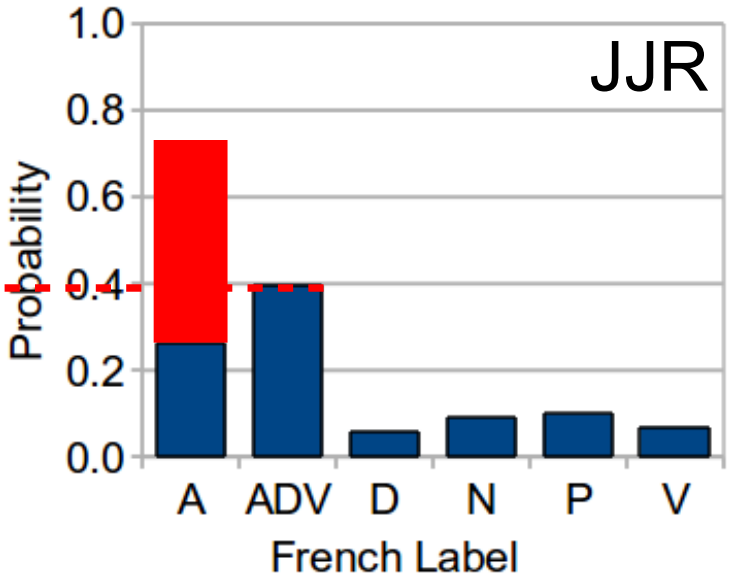


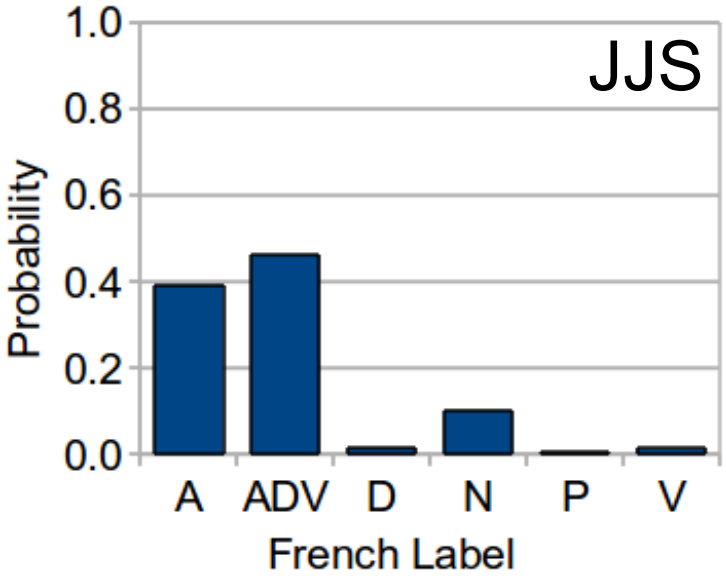




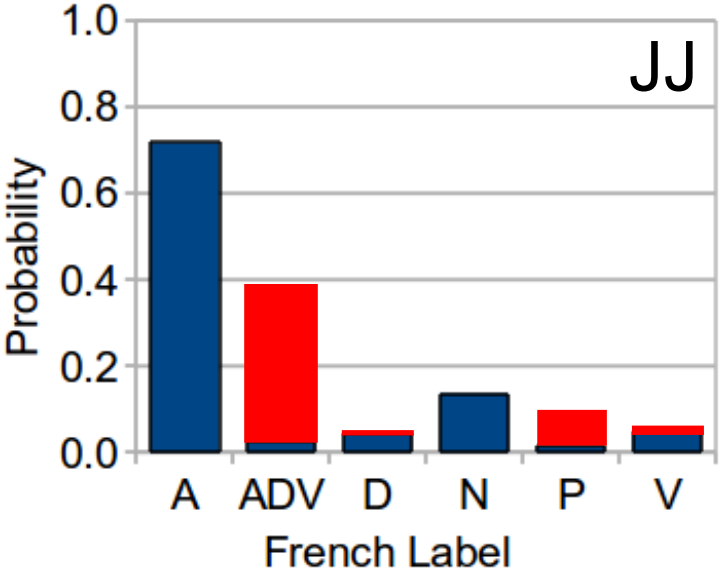
L_1 Alignment Distance

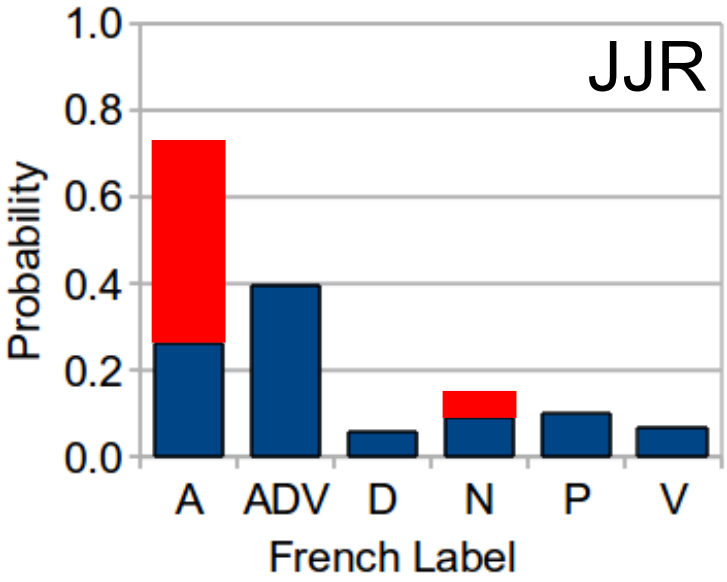


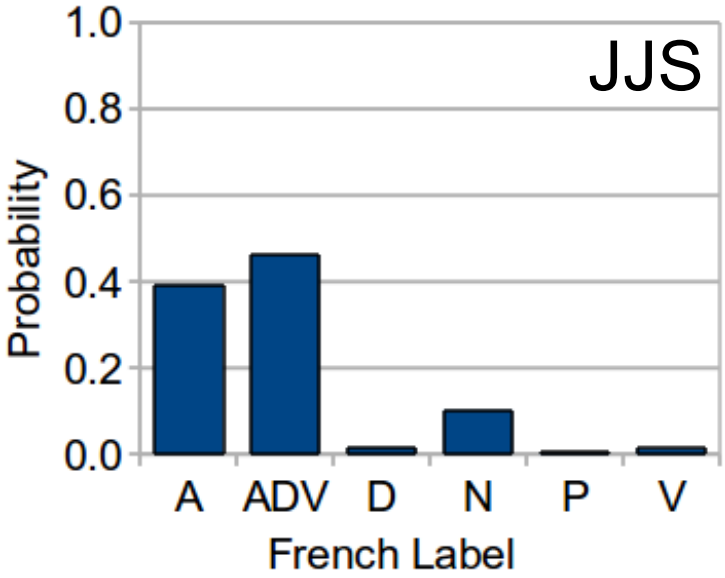




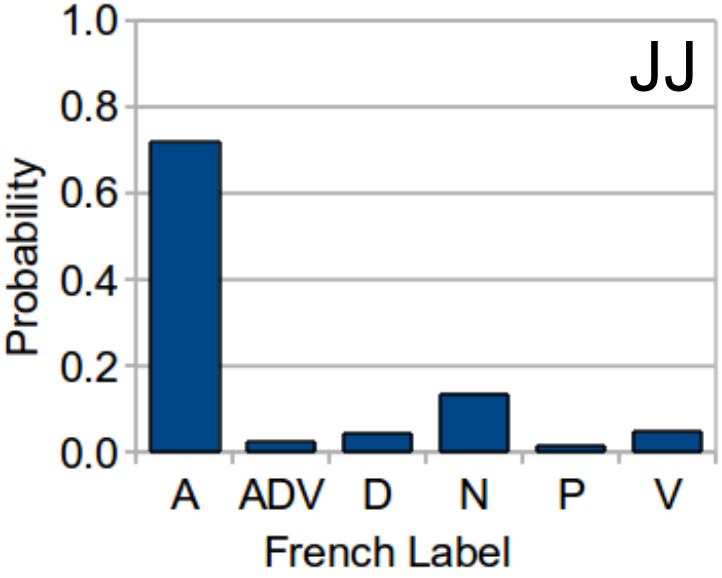
L_1 Alignment Distance



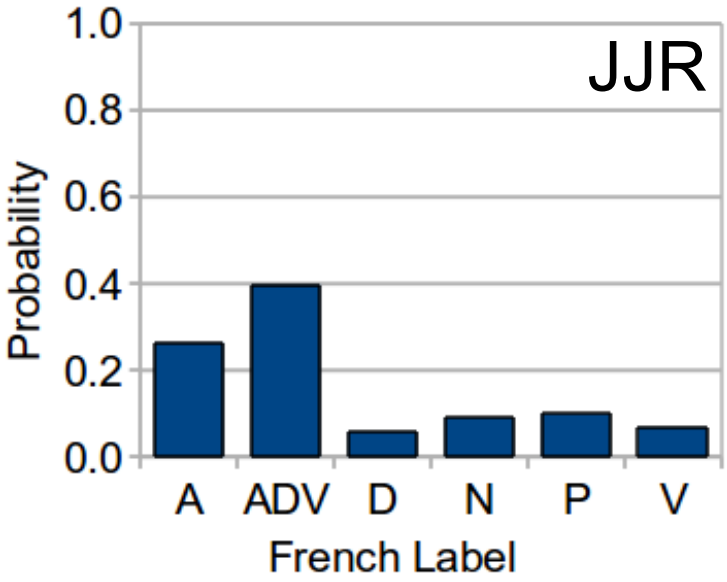




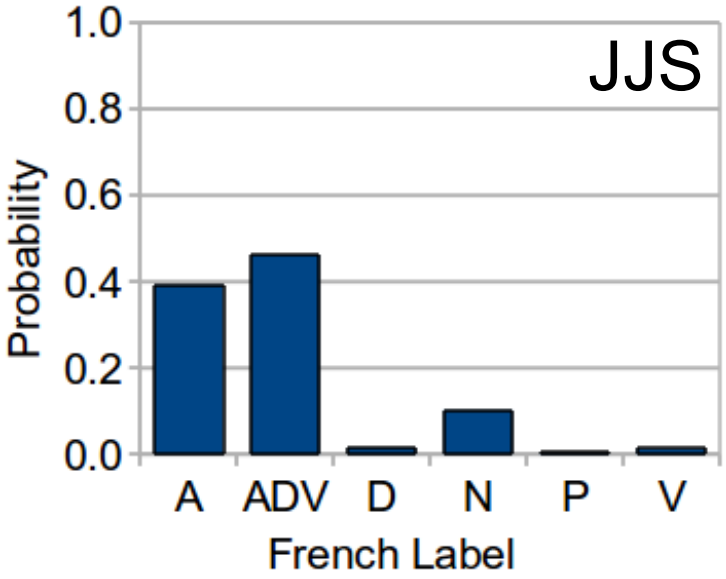
L_1 Alignment Distance



0.9941



0.8730



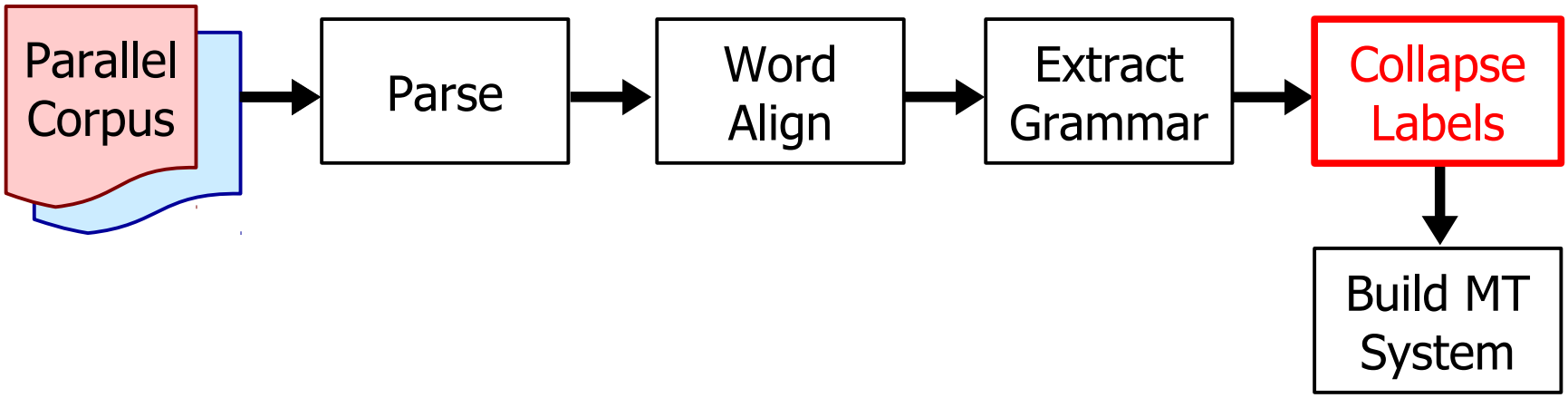
0.3996

Label Collapsing Algorithm

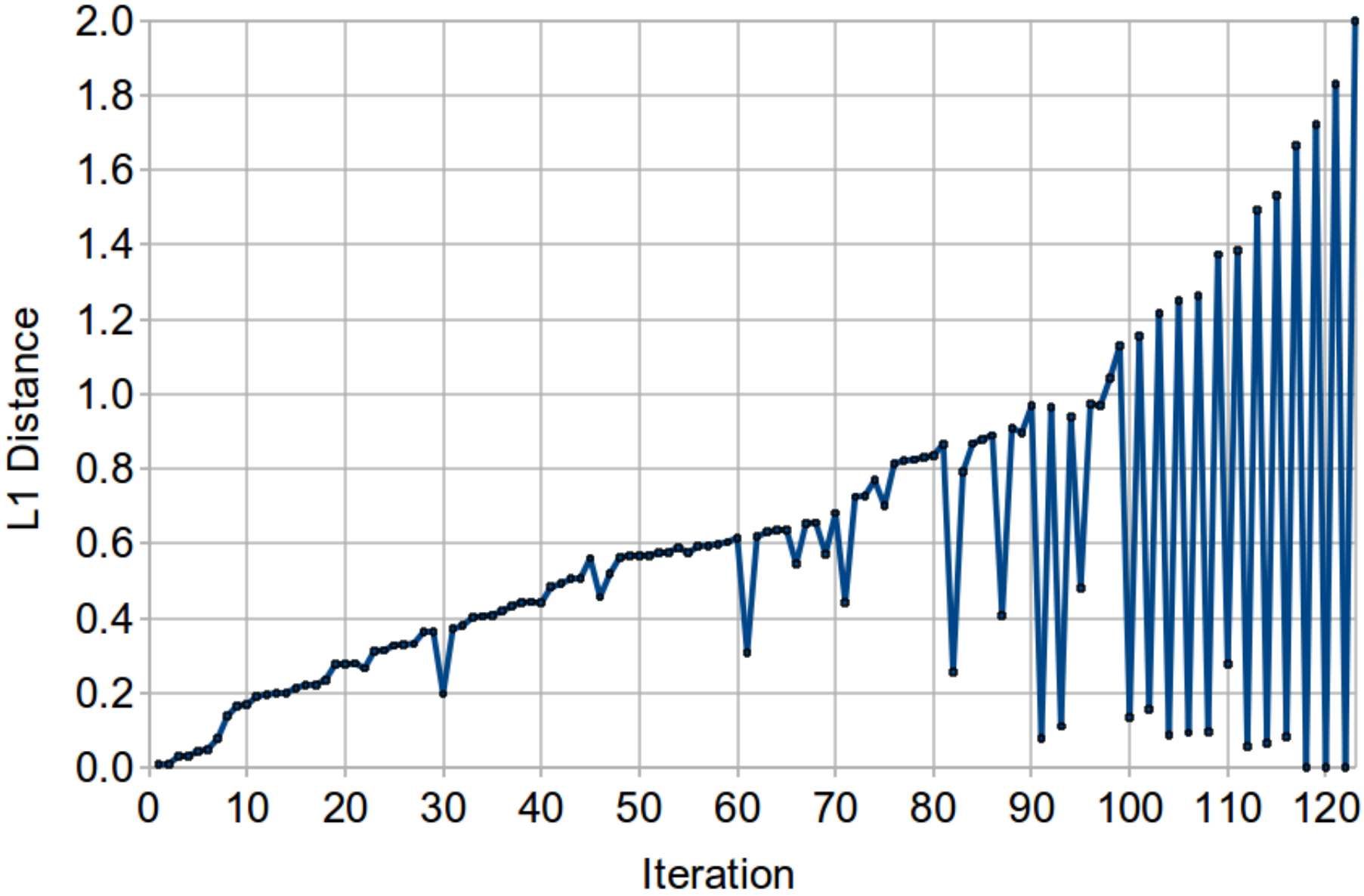
- Extract baseline grammar from aligned tree pairs (e.g. Lavie et al. [2008])
- Compute label alignment distributions
- Repeat until stopping point:
  - Compute L_1 distance between all pairs of source and target labels
  - Merge the label pair with smallest distance
  - Update label alignment distributions

Experiment 1

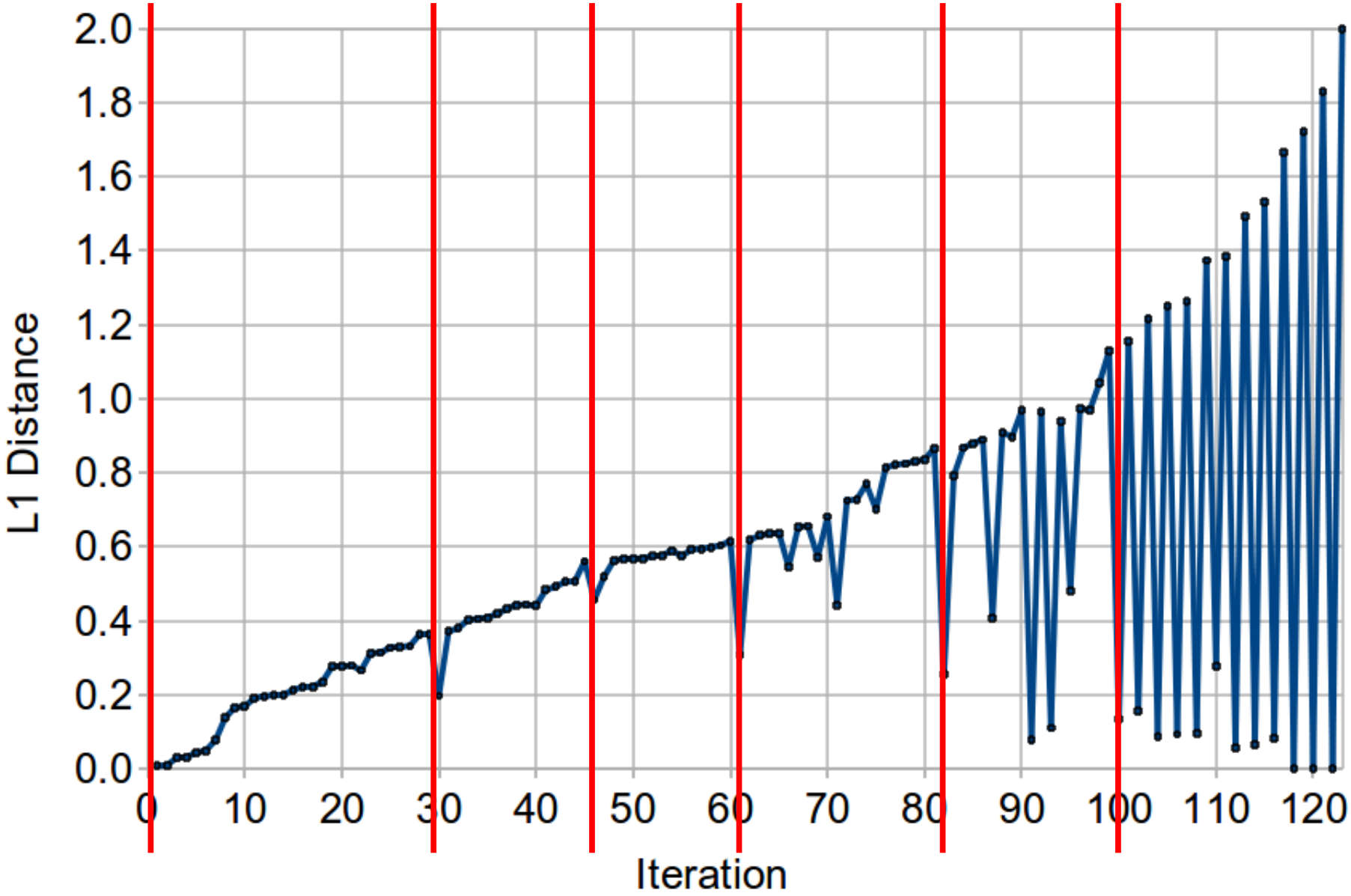
- Goal: Explore effect of collapsing with respect to stopping point
- Data: Chinese–English FBIS corpus (302 k)



Experiment 1



Experiment 1



Effect on Label Set

- Number of unique labels in grammar

	Zh	En	Joint
Baseline	55	71	1556
Iter. 29	46	51	1035
Iter. 45	38	44	755
Iter. 60	33	34	558
Iter. 81	24	22	283
Iter. 99	14	14	106

Effect on Grammar

- Split grammar into three partitions:
  - Phrase pair rules

$$\text{NN::NN} \rightarrow [\text{友好}]::[\text{friendship}]$$

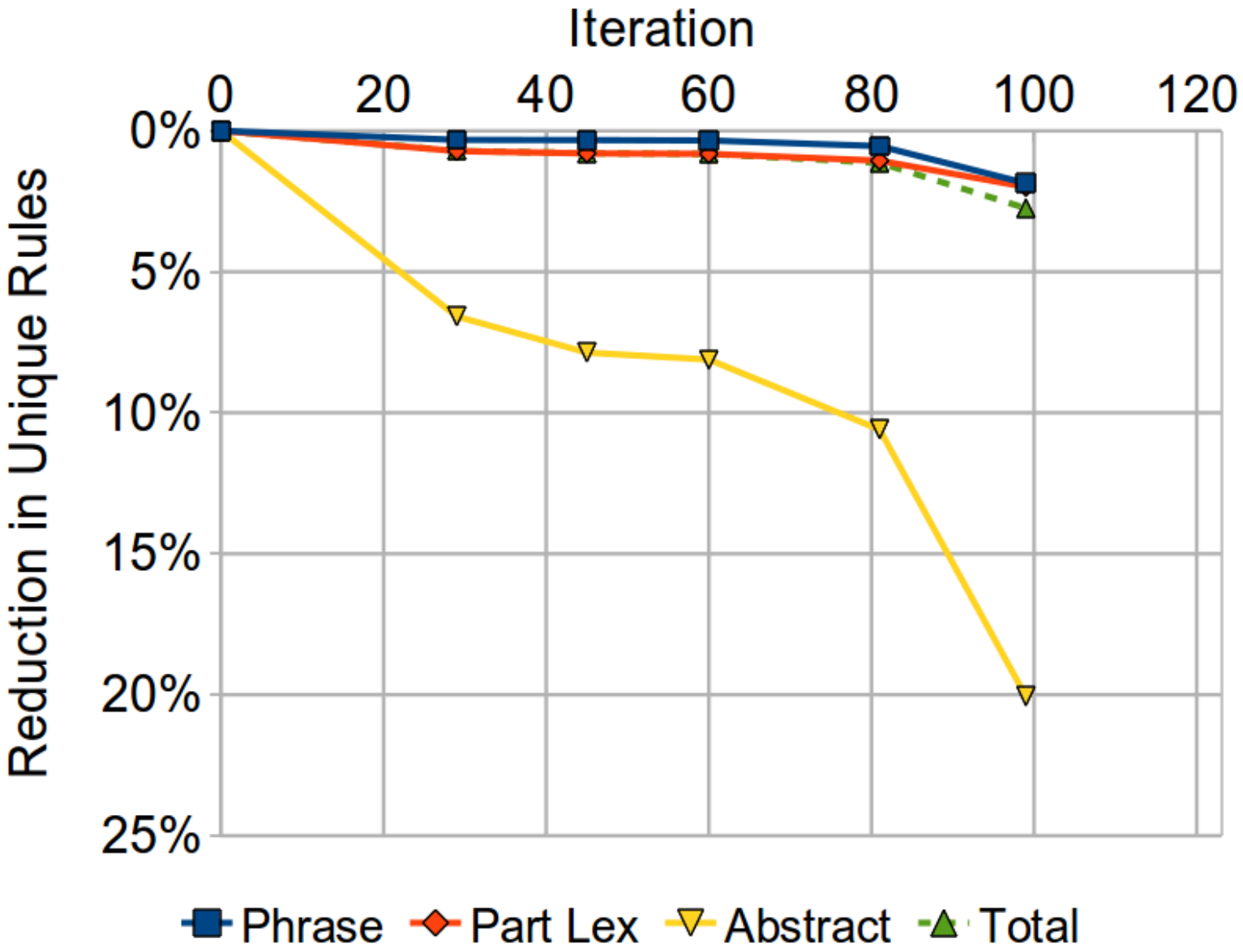
  - Partially lexicalized grammar rules

$$\text{NP::NP} \rightarrow [2000\ \text{年}\ \ \text{NN}^1]::[\text{the 2000 NN}^1]$$

  - Fully abstract grammar rules

$$\text{VP::ADJP} \rightarrow [\text{VV}^1\ \text{VV}^2]::[\text{RB}^1\ \text{VBN}^2]$$

Effect on Grammar



Effect on Metric Scores

- NIST MT '03 Chinese–English test set
- Results averaged over four tune/test runs

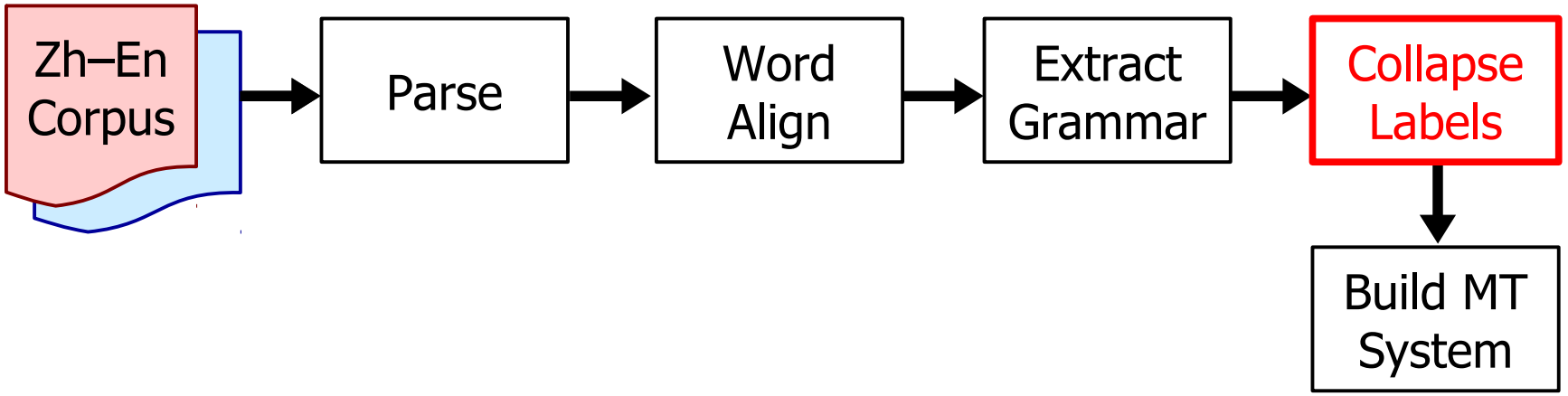
	BLEU	METR	TER
Baseline	24.43	54.77	68.02
Iter. 29	27.31	55.27	63.24
Iter. 45	27.10	55.24	63.41
Iter. 60	27.52	55.32	62.67
Iter. 81	26.31	54.63	63.53
Iter. 99	25.89	54.76	64.82

Effect on Decoding

- Different outputs produced
  - Collapsed 1-best in baseline 100-best: 3.5%
  - Baseline 1-best in collapsed 100-best: 5.0%
- Different hypergraph entries explored in cube pruning
  - 90% of collapsed entries not in baseline
  - Overlapping entries tend to be short
- Hypothesis: different rule possibilities lead search in complementary direction

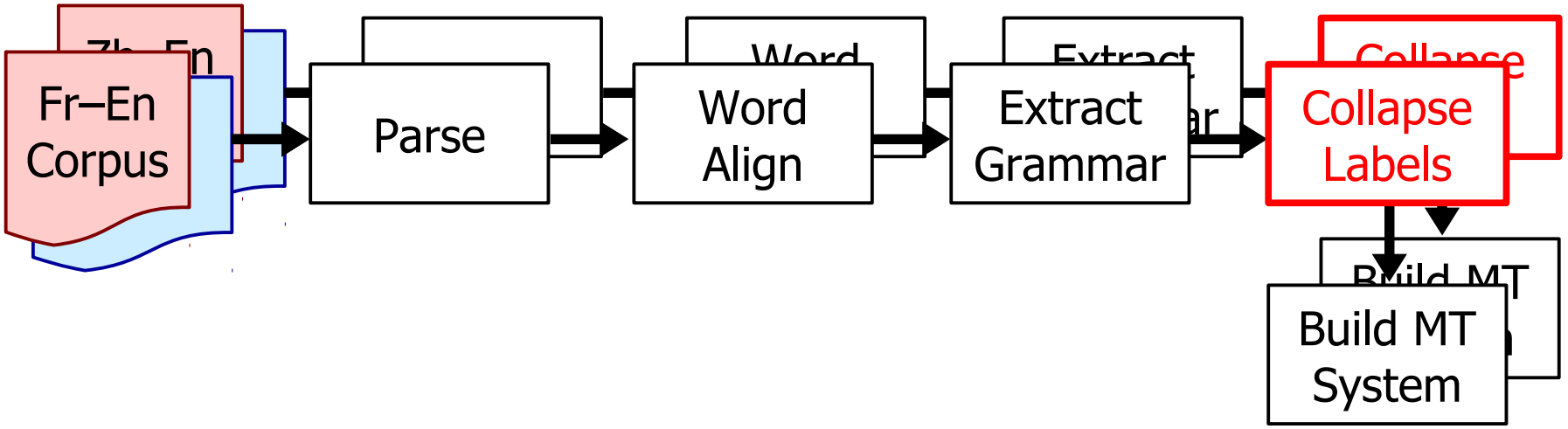
Experiment 2

- Goal: Explore effect of collapsing across language pairs
- Data: Chinese–English FBIS corpus, French–English WMT 2010 data



Experiment 2

- Goal: Explore effect of collapsing across language pairs
- Data: Chinese–English FBIS corpus,
  French–English WMT 2010 data



Effect on English Collapsing

- Adverbs
  - Zh–En: RB, RBR
  - Fr–En: RBR, RBS
- Verbs
  - Zh–En: VB, VBG, VBN
  - Fr–En: VB, VBD, VBN, VBP, VBZ, MD
- *Wh*-phrases
  - Zh–En: ADJP, WHADJP; ADVP, WHADVP
  - Fr–En: PP, WHPP

Effect on Label Set

- Full subtype collapsing

VNV VSB VRD VPT VCD VCP

VC

- Partial subtype collapsing

NN NNS NNPS NNP

N

- Combination by syntactic function

RRC WHADJP INTJ

INS

Conclusions

- Can effectively coarsen labels based on alignment distributions
- Significantly improved metric scores at all attempted stopping points
- Reduces rule sparsity more than labeling ambiguity
- Points decoder in different direction
- Different results for different language pairs or grammars

Future Work

- Take rule context into account

$[NP::NP] \rightarrow [D^1\ N^2]::[DT^1\ NN^2]$ la voiture / the car

$[NP::NP] \rightarrow [\text{les}\ N^2]::[NNS^2]$ les voitures / cars

- Try finer-grained label sets [Petrov et al. 2006]

NP	NP-0, NP-1, ..., NP-30
VBN	VBN-0, VBN-1, ..., VBN-25
RBS	RBS-0

- Non-greedy collapsing

References

- Chiang (2010), "Learning to translate with source and target syntax," ACL
- Lavie, Parlikar, and Ambati (2008), "Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora," SSST-2
- Petrov, Barrett, Thibaux, and Klein (2006), "Learning accurate, compact, and interpretable tree annotation," ACL/COLING
- Venugopal, Zollmann, Smith, and Vogel (2009), "Preference grammars: Softening syntactic constraints to improve statistical machine translation," NAACL