Statistical MT with Syntax and Morphology: Challenges and Some Solutions

Alon Lavie

LTI Colloquium
September 2, 2011

Joint work with:
Greg Hanneman, Jonathan Clark, Michael Denkowski, Kenneth Heafield, Hassan Al-Haj, Michelle Burroughs, Silja Hildebrand and Nizar Habash

Outline

- Morphology, Syntax and the challenges they pose for MT
- Frameworks for Statistical MT with Syntax and Morphology
- Impact of Morphological Segmentation on large-scale Phrase-based SMT of English to Arabic
- Learning Large-scale Syntax-based Translation Models from Parsed Parallel Corpora
- Statistical MT between Hebrew and Arabic
- Improving Category Labels for Syntax-based SMT
- Conclusions

Phrase-based SMT

- **Acquisition:** Learn bilingual correspondences for word and multi-word sequences from large volumes of sentence-level parallel corpora
- **Decoding:** Efficiently search a (sub) space of possible combinations of phrase translations that generate a complete translation for the given input
  - Limited reordering of phrases
  - Linear model for combining the collection of feature scores (translation model probabilities, language model, other), optimized on tuning data

Phrase-based SMT

- **Strengths:**
  - **Simple** (naïve) modeling of the language translation problem!
  - Acquisition requires just raw sentence-aligned parallel data and monolingual data for language modeling – plenty around! And constantly growing (for some language pairs…)
  - Works surprisingly well – for some language pairs!
- **Weaknesses:**
  - Simple **(naïve)** modeling of the language translation problem!
  - Cannot model and generate the correct translation for many linguistic phenomena across languages – both common and rare!
  - Doesn't generalize well – models are purely lexical
  - Performance varies widely across language pairs and domains
  - These issues are particularly severe for languages with rich morphology and languages with highly-divergent syntax and semantics

Challenges: Morphology

- Most languages have far richer and more complex word morphology than English
- Example: Hebrew
  - וכשתפתח את הקובץ החדש
  - the new the file ET and when you (will) open
  - and when you open the new file
- Challenges for Phrase-based Statistical MT:
  - Data sparsity in acquisition and decoding
    - Many forms of related words (i.e. inflected forms of verbs) seen only a few times in the parallel training data
    - Many forms not seen at all – unknown words during decoding
  - Difficulty in acquiring accurate one-to-many word-alignment mappings
  - Complex cross-lingual mappings of morphology and syntax
- Non-trivial solution: morphological segmentation and/or deep analysis

Morphology within SMT Frameworks

- **Options for handling morphology:**
  - Morpheme segmentation, possibly including some mapping into base or canonical forms
  - Full morphological analysis, with a detailed structural representation, possibly including the extraction of subset of features for MT
- What types of analyzers are available? How accurate?
- How do they deal with morphological ambiguity?
- Computational burden of analyzing massive amounts of training data and running analyzer during decoding
- What should we segment and what features to extract for best MT performance?
- Impact on language modeling

Challenges: Syntax

- Syntax of the source language is different than syntax of the target language:
  - Word order within constituents:
    - English NPs: art adj n the big boy
    - Hebrew NPs: art n art adj ha-yeled ha-gadol הילד הגדול
  - Constituent structure:
    - English is SVO: Subj Verb Obj I saw the man
    - Modern Standard Arabic is (mostly) VSO: Verb Subj Obj
  - Different verb syntax:
    - Verb complexes in English vs. in German

      I can eat the apple *Ich kann den apfel essen*
  - Case marking and free constituent order
    - German and other languages that mark case:

      den apfel esse Ich *the*(acc) *apple eat I*(nom)

Challenges: Syntax

- Challenges of divergent syntax on Statistical MT:
  - Lack of abstraction and generalization:
    - [ha-yeled ha-gadol] ➔ [the big boy]
    - [ha-yeled] + [ha-katan] ➔ [the boy] + [the small]
    - Desireable: art n art adj ➔ art adj n
    - Requires deeper linguistic annotation of the training data and appropriately-abstract translations models and decoding algorithms
  - Long-range reordering of syntactic structures:
    - Desireable translation rule for Arabic to English:

      V NP_Subj NP_Obj ➔ NP_Subj V NP_Obj
    - Requires identifying the appropriate syntactic structure on the source language and acquiring rules/models of how to correctly map them into the target language
    - Requires deeper linguistic annotation of the training data and appropriately-abstract translations models and decoding algorithms

Syntax-Based SMT Models

- Various proposed models and frameworks, no clear winning consensus model as of yet
- Models represent pieces of hierarchical syntactic structure on source and target languages and how they map and combine
- Most common representation model is *Synchronous Context-Free Grammar* (S-CFGs), often augmented with statistical features
  - NP::NP → [Det$_1$ N$_2$ Det$_1$ Adj$_3$]::[Det$_1$ Adj$_3$ N$_2$]
- How are these models acquired?
  - **Supervised:** acquired from parallel-corpora that are annotated in advance with syntactic analyses (parse trees) for each sentence
    - Parse source language, target language or both?
    - Computational burden of parsing all the training data
    - Parsing ambiguity
    - What syntactic labels should be used?
  - **Unsupervised:** induce the hierarchical structure and source-target mappings directly from the raw parallel data

What This Talk is About

- Research work within my group and our collaborators addressing some specific instances of such MT challenges related to morphology and syntax

1. Impact of Arabic morphological segmentation on broad-scale English-to-Arabic Phrase-based SMT
2. Learning of syntax-based synchronous context-free grammars from vast parsed parallel corpora
3. Exploring the Category Label Granularity Problem in Syntax-based SMT

The Impact of Arabic Morphological Segmentation on Broad-Scale Phrase-based SMT

Joint work with Hassan Al-Haj

with contributions from Nizar Habash,

Kenneth Heafield, Silja Hildebrand and Michael Denkowski

lti

Carnegie Mellon

Motivation

- Morphological segmentation and tokenization decisions are important in phrase-based SMT
  - Especially for morphologically-rich languages
- Decisions impact the entire pipeline of training and decoding components
- Impact of these decisions is often difficult to predict in advance
- **Goal:** a detailed investigation of this issue in the context of phrase-based SMT between English and Arabic
  - Focus on segmentation/tokenization of the Arabic (not English)
  - Focus on translation from English into Arabic

Research Questions

- Do Arabic segmentation/tokenization decisions make a significant difference even in large training data scenarios?
- English-to-Arabic vs. Arabic-to-English
- What works best and why?
- Additional considerations or impacts when translating into Arabic (due to detokenization)
- Output Variation and Potential for System Combination?

Methodology

- Common large-scale training data scenario (NIST MT 2009 English-Arabic)
- Build a rich spectrum of Arabic segmentation schemes (nine different schemes)
  - Based on common detailed morphological analysis using MADA (Habash et al.)
- Train nine different complete end-to-end English-to-Arabic (and Arabic-to-English) phrase-based SMT systems using Moses (Koehn et al.)
- Compare and analyze performance differences

Arabic Morphology

- Rich inflectional morphology with several classes of clitics and affixes that attach to the word
- conj + part + art + base + pron

CONJ	w+ (*and*), f+ (*then*)
PART	l+ (*to/for*), b+ (*by/with*), k+ (*as/such*) s+ *will/future*.
DET	Al+(*the*)
PRON	+h (+O:3MS, +P:3MS) +hA (+O:3FS,+P:3FS) +hm (+O:3MP,+P:3MP) +hmA (+O:3D,+P:3D) +hn (+O:3FP, +P:3FP) +k (+O:2FS,+P:2FS,+O:2MS,+P:2MS) +km (+O:2MP,+P:2MP) +kmA (+O:2D,+P:2D) +kn (+O:2FP,+P:2FP) +nA (+O:1P,+P:1P) +y (+O:1S,+P:1S)

Table 1. Arabic clitics divided to 4 classes.

Arabic Orthography

- Deficient (and sometimes inconsistent) orthography
  - Deletion of short vowels and most diacritics
  - Inconsistent use of ŢÇ¥Ч
  - Inconsistent use of ,
- Common Treatment (Arabic→English)
  - Normalize the inconsistent forms by collapsing them
- Clearly undesirable for MT into Arabic
  - Enrich: use MADA to disambiguate and produce the full form
  - Correct full-forms enforced in training, decoding and evaluation

Arabic Segmentation and Tokenization Schemes

- Based on common morphological analysis by MADA and tokenization byTOKAN (Habash et el.)
- Explored nine schemes (coarse to fine):
  - UT: unsegmented (full enriched form)
  - S0: w + REST
  - S1: w|f + REST
  - S2: w|f + part|art + REST
  - S3: w|f + part/s|art + base + pron-MF
  - S4: w|f + part|art + base + pron-MF
  - S4SF: w|f + part|art + base + pron-SF
  - S5: w|f + part +art + base + pron-MF
  - S5SF: w|f + part + art + base + pron-SF

Arabic Segmentation and Tokenization Schemes

- Based on common morphological analysis by MADA and tokenization by TOKAN (Habash et el.)
- Explored nine schemes (coarse to fine):
  - UT: unsegmented (full enriched form)
  - S0: w + REST
  - S1: w|f + REST
  - S2: w|f + part|art + REST
  - **S3:** w|f + part/s|art + base + **pron-MF**
  - **S4:** w|f + part|art + base + **pron-MF**
  - S4SF: w|f + part|art + base + pron-SF
  - **S5:** w|f + part +art + base + **pron-MF**
  - S5SF: w|f + part + art + base + pron-SF

Morphological Forms!

Arabic Segmentation and Tokenization Schemes

- Based on common morphological analysis by MADA and tokenization byTOKAN (Habash et el.)
- Explored nine schemes (coarse to fine):
  - UT: unsegmented (full enriched form)
  - S0: w + REST
  - S1: w|f + REST
  - S2: w|f + part|art + REST
  - S3: w|f + part/s|art + base + pron-MF
  - S4: w|f + part|art + base + pron-MF
  - **S4SF:** w|f + part|art + base + **pron-SF**
  - S5: w|f + part +art + base + pron-MF
  - **S5SF:** w|f + part + art + base + **pron-SF**

Surface Forms!

Arabic Segmentation and Tokenization Schemes

- Based on common morphological analysis by MADA and tokenization byTOKAN (Habash et el.)
- Explored nine schemes (coarse to fine):
  - UT: unsegmented (full enriched form)
  - S0: w + REST
  - S1: w|f + REST
  - S2: w|f + part|art + REST
  - **S3: w|f + part/s|art + base + pron-MF** **Original PATB**
  - **S4: w|f + part|art + base + pron-MF** **ATBv3**
  - S4SF: w|f + part|art + base + pron-SF
  - S5: w|f + part + art + base + pron-MF
  - S5SF: w|f + part + art + base + pron-SF

Arabic Segmentation Schemes

Input	wbAlnsbp lAyTAlyA fAnh yEny AnhA sttSrf kdwlp Sgyrp ttxlY En ms&wlyAthA
Gloss	and regarding to italy this means that it will act as a country small giving up its responsibilities
English	And regarding Italy, this mean that it will act as a small country giving up its responsibilities
UT	wbAlnsbp l<yTAlyA f>nh yEny >nhA sttSrf kdwlp Sgyrp ttxlY En ms&wlyAthA
S0	w+ bAlnsbp l<yTAlyA f>nh yEny >nhA sttSrf kdwlp Sgyrp ttxlY En ms&wlyAthA
S1	w+ bAlnsbp l<yTAlyA f+ >nh yEny >nhA sttSrf kdwlp Sgyrp ttxlY En ms&wlyAthA
S2	w+ b+ Alnsbp l+ <yTAlyA f+ >nh yEny >nhA s+ ttSrf k+ dwlp Sgyrp ttxlY En ms&wlyAthA
S3	w+ b+ Alnsbp l+ <yTAlyA f+ >n +O:3MS yEny >n +O:3FS sttSrf k+ dwlp Sgyrp ttxlY En ms&wlyAt +P:3FS
S4	w+ b+ Alnsbp l+ <yTAlyA f+ >n +O:3MS yEny >n +O:3FS s+ ttSrf k+ dwlp Sgyrp ttxlY En ms&wlyAt +P:3FS
S5	w+ b+ Al+ nsbp l+ <yTAlyA f+ >n +O:3MS yEny >n +O:3FS s+ ttSrf k+ dwlp Sgyrp ttxlY En ms&wlyAt +P:3FS
S5SF	w+ b+ Al+ nsbp l+ <yTAlyA f+ >n +h yEny >n +hA s+ ttSrf k+ dwlp Sgyrp ttxlY En ms&wlyAt +hA

Table 2. The different tokenization schemes exemplified on the same sentence.

S	Token#	Type #	OOV#
UT	136,280,410	653,584	85
S0	145,826,275	566,024	76
S1	146,162,567	552,150	76
S2	154,974,999	475,335	68
S3	160,194,619	425,645	62
S4	160,599,031	418,832	62
S5	199,179,300	391,190	59

Table 3. tokens, and types count of the Arabic side of the training data for the different schemes and the out-of-vocabulary tokens on NIST MT02 test set.

46% more tokens

40% fewer types

Reduction in OOVs

Previous Work

- Most previous work has looked at these choices in context of Arabic→English MT
  - Most common approach is to use PATB or ATBv3
- (Badr et al. 2006) investigated segmentation impact in the context of English→Arabic
  - Much smaller-scale training data
  - Only a small subset of our schemes

Arabic Detokenization

- English-to-Arabic MT system trained on segmented Arabic forms will decode into segmented Arabic
  - Need to put back together into full form words
  - Non-trivial because mapping isn't simple concatenation and not always one-to-one
  - Detokenization can introduce errors
  - The more segmented the scheme, the more potential errors in detokenization

Arabic Detokenization

- We experimented with several detokenization methods:
  - C: simple concatenation
  - R: List of detokenization rules (Badr et al. 2006)
  - T: Mapping table constructed from training data (with likelihoods)
  - T+C: Table method with backoff to C
  - T+R: Table method with backoff to R
  - T+R+LM: T+R method augmented with a 5-gram LM of full-forms and viterbi search for max likelihood sequence.
- T+R was the selected approach for this work

Experimental Setup

- NIST MT 2009 constrained training parallel-data for Arabic-English:
  - ~5 million sentence-pairs
  - ~150 million unsegmented Arabic words
  - ~172 million unsegmented English words
- Preprocessing:
  - English tokenized using Stanford tokenizer and lower-cased
  - Arabic analyzed by MADA, then tokenized using scripts and TOKAN according to the nine schemes
- Data Filtering: sentence pairs with > 99 tokens on either side or ratio of more than 4-to-1 were filtered out

Training and Testing Setup

- Standard training pipeline using Moses
  - Word Alignment of tokenized data using MGIZA++
  - Symetrized using grow-diag-final-and
  - Phrase extraction with max phrase length 7
  - Lexically conditioned distortion model conditioned on both sides
- Language Model: 5-gram SRI-LM trained on tokenized Arabic-side of parallel data (152 million words)
  - Also trained 7-gram LM for S4 and S5
- Tune: MERT to BLEU-4 on MT-02
- Decode with Moses on MT-03, MT-04 and MT-05
- Detokenized with T+R method
- Scored using BLEU, TER and METEOR on detokenized output

English-to-Arabic Results

System	BLEU	TER	METEOR
UT	35.66	50.76	51.21
S0	**36.25**	**50.98**	**51.60**
S1	35.74	51.47	50.98
S2	*35.05*	*53.16*	*49.81*
S3	36.19	50.49	51.75
S4	**36.22**	**50.61**	**51.58**
S5	34.93	51.77	49.96
S4SF	35.83	50.88	51.48
S5SF	*33.64*	*52.73*	*48.90*
S4,7gram	35.81	50.92	51.26
S5,7gram	34.84	51.88	50.10

MT03

System	BLEU	TER	METEOR
UT	31.53	56.15	45.55
S0	**31.80**	**56.26**	**45.87**
S1	31.46	57.08	45.17
S2	*29.89*	*59.49*	*44.03*
S3	31.73	56.25	45.81
S4	**31.90**	**55.86**	**45.90**
S5	30.87	57.56	44.52
S4SF	31.99	55.90	45.84
S5SF	*30.06*	*57.83*	*43.67*
S4,7gram	31.46	56.04	45.60
S5,7gram	30.91	57.31	44.47

MT04

System	BLEU	TER	METEOR
UT	38.40	47.94	53.96
S0	**38.83**	**48.42**	**54.13**
S1	38.29	48.84	53.40
S2	*37.29*	*51.00*	*52.72*
S3	38.55	48.22	54.33
S4	**38.55**	**48.01**	**54.21**
S5	37.72	49.65	52.94
S4SF	38.15	48.28	54.01
S5SF	*36.80*	*49.91*	*52.00*
S4,7gram	38.32	48.19	54.07
S5,7gram	37.72	49.23	52.81

MT05

Analysis

- **Complex picture:**
  - Some decompositions help, others don't help or even hurt performance
- Segmentation decisions really matter – even with large amounts of training data:
  - Difference between best (S0) and worst (S5SF)
    - **On MT03 : +2.6 BLEU, -1.75 TER, +2.7 METEOR points**
- Map Key Reminder:
  - S0: w+REST, S2: conj+part|art+REST, S4: (ATBv3) split all except for the art, S5: split everything (pron in morph. form)
- S0 and S4 consistently perform the best, are about equal
- S2 and S5 consistently perform the worst
- S4SF and S5SF usually worse than S4 and S5

Analysis

- Simple decomposition S0 (just the “w” conj) works as well as any deeper decomposition
- S4 (ATBv3) works well also for MT into Arabic
- Decomposing the Arabic definite article consistently hurts performance
- Decomposing the prefix particles sometimes hurts
- Decomposing the pronominal suffixes (MF or SF) consistently helps performance
- 7-gram LM does not appear to help compensate for fragmented S4 and S5

Analysis

- Clear evidence that splitting off the Arabic definite article is bad for English→Arabic
  - S4→S5 results in 22% increase in PT size
  - Significant increase in translation ambiguity for short phrases
  - Inhibits extraction of some longer phrases
  - Allows ungrammatical phrases to be generated:
    - Middle East → **Al**$rq **Al**>wsT
    - Middle East → $rq >qsT
    - Middle East → $rq **Al**>wsT

Output Variation

- How different are the translation outputs from these MT system variants?
  - Upper-bound: Oracle Combination on the single-best hypotheses from the different systems
    - Select the best scoring output from the nine variants (based on posterior scoring against the reference)
  - Work in Progress - actual system combination:
    - Hypothesis Selection
    - CMU Multi-Engine MT approach
    - MBR

Oracle Combination

MT03

System	BLEU	TER	METEOR
Best Ind. (S0)	36.25	50.98	51.60
Oracle Combination	**41.98**	**44.59**	**58.36**

MT04

System	BLEU	TER	METEOR
Best Ind. (S4)	31.90	55.86	45.90
Oracle Combination	**37.38**	**50.34**	**52.61**

MT05

System	BLEU	TER	METEOR
Best Ind. (S0)	38.83	48.42	54.13
Oracle Combination	**45.20**	**42.14**	**61.24**

Output Variation

- Oracle gains of 5-7 BLEU points from selecting among nine variant hypotheses
  - Very significant variation in output!
  - Better than what we typically see from oracle selections over large n-best lists (for n=1000)

Arabic-to-English Results

	BLEU	TER	METEOR
UT	49.55	42.82	72.72
S0	49.27	43.23	72.26
S1	*49.17*	*43.03*	*72.37*
S2	49.97	42.82	73.15
S3	49.15	43.16	72.49
S4	49.70	42.87	72.99
S5	**50.61**	**43.17**	**73.16**
S4SF	49.60	43.53	72.57
S5SF	49.91	43.00	72.62

MT03

Analysis

- Still some significant differences between the system variants
  - Less pronounced than for English→Arabic
- Segmentation schemes that work best are different than in the English→Arabic direction
- S4 (ATBv3) works well, but isn't the best
- More fragmented segmentations appear to work better
- Segmenting the Arabic definite article is no longer a problem
  - S5 works well now
- We can leverage from the output variation
  - Preliminary hypothesis selection experiments show nice gains

Conclusions

- Arabic segmentation schemes has a significant impact on system performance, even in very large training data settings
  - Differences of 1.8-2.6 BLEU between system variants
- Complex picture of which morphological segmentations are helpful and which hurt performance
  - Picture is different in the two translation directions
  - Simple schemes work well for English→Arabic, less so for Arabic→English
  - Splitting off Arabic definite article hurts for English→Arabic
- Significant variation in the output of the system variants can be leveraged for system combination

A General-Purpose Rule Extractor for SCFG-Based Machine Translation

Joint work with

Greg Hanneman and **Michelle Burroughs**

lti

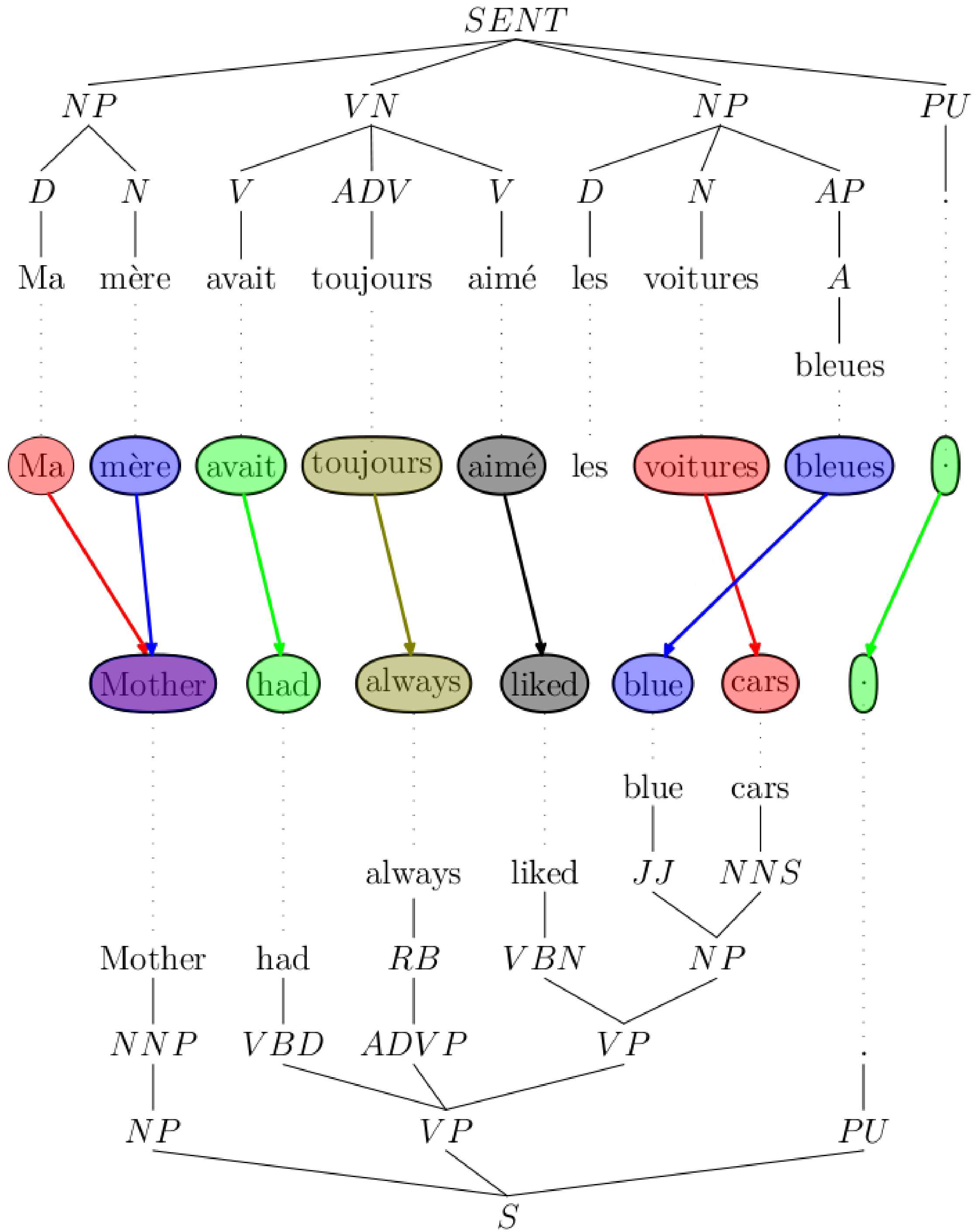
Carnegie Mellon

S-CFG Grammar Extraction

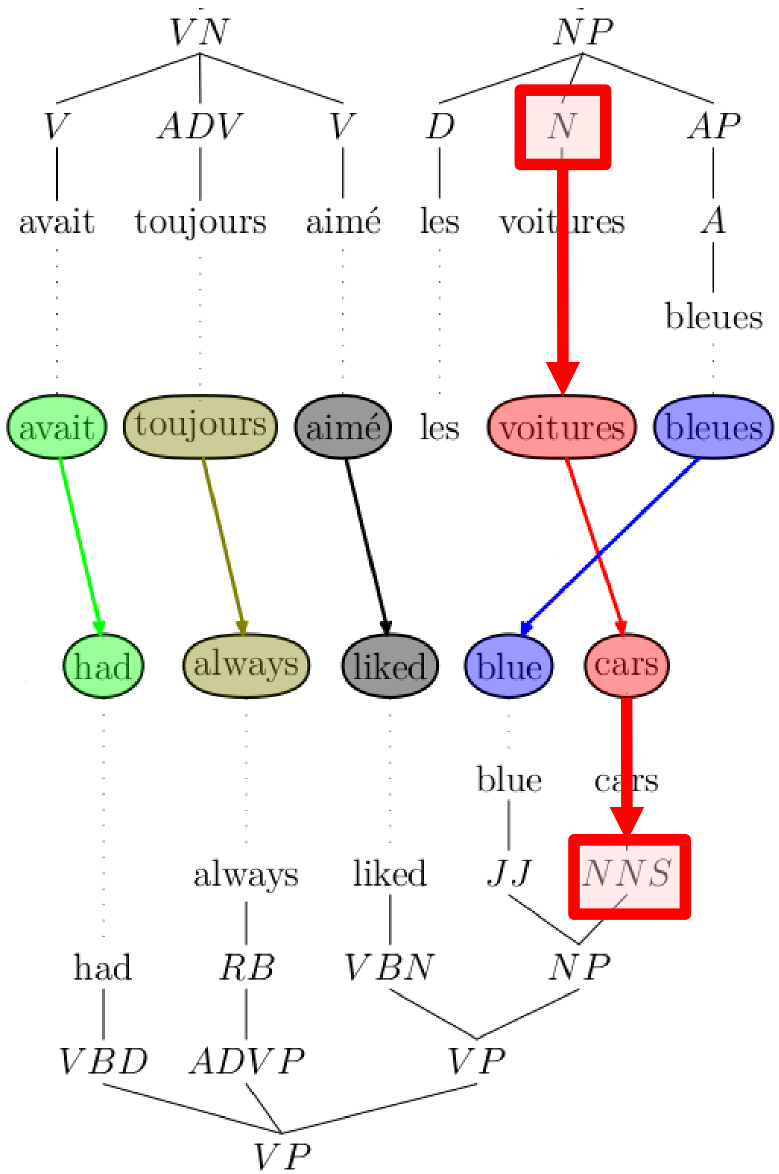
- Inputs:
  - Word-aligned sentence pair
  - Constituency parse trees on one or both sides
- Outputs:
  - Set of S-CFG rules derivable from the inputs, possibly according to some constraints
- Implemented by:
  - Hiero [Chiang 2005] GHKM [Galley et al. 2004]
    Chiang [2010] Stat-XFER [Lavie et al. 2008]
    SAMT [Zollmann and Venugopal 2006]

S-CFG Grammar Extraction

- **Our goals:**
  - Support for two-side parse trees by default
  - Extract greatest number of syntactic rules…
  - Without violating constituent boundaries
- **Achieved with:**
  - Multiple node alignments
  - Virtual nodes
  - Multiple right-hand-side decompositions
- First grammar extractor to do all three

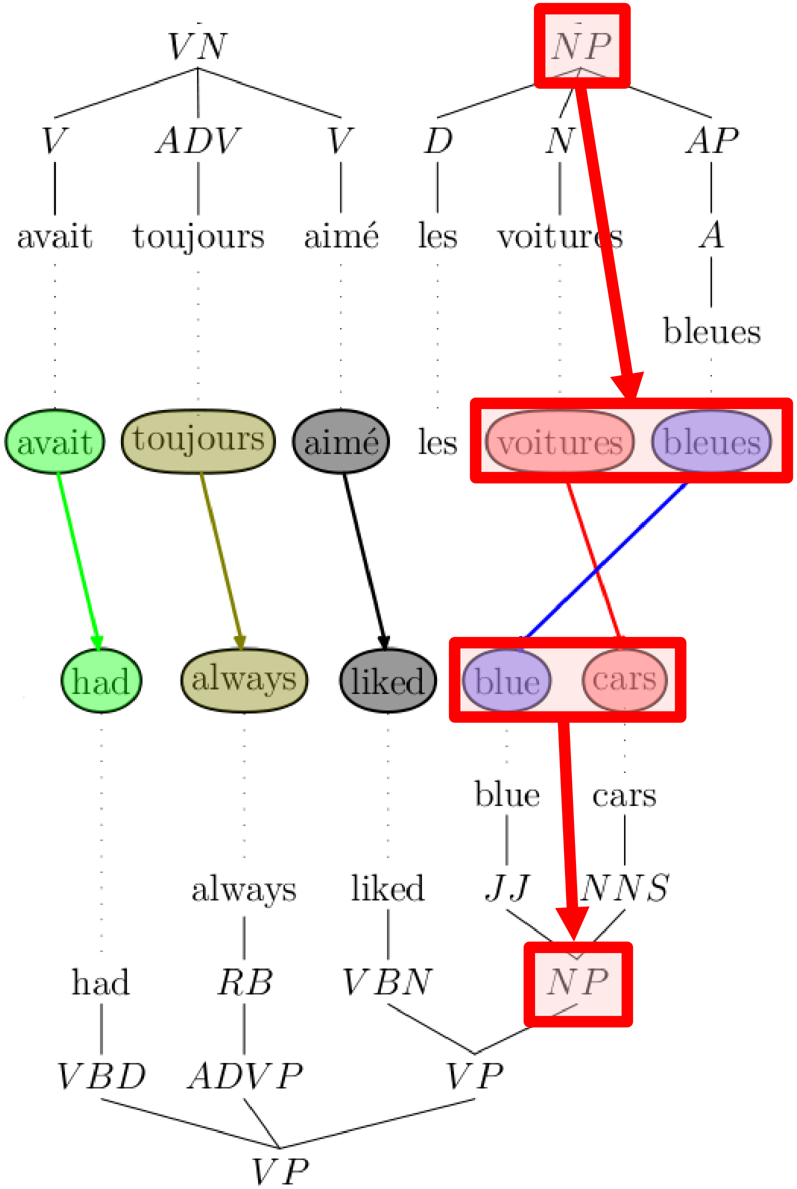

SENT
NP
VN
NP
PU
D
N
V
ADV
V
D
N
AP
.
Ma
mère
avait
toujours
aimé
les
voitures
A
bleues
Ma
mère
avait
toujours
aimé
les
voitures
bleues
.
Mother
had
always
liked
blue
cars
.
blue
cars
always
liked
JJ
NNS
Mother
had
RB
VBN
NP
NNP
VBD
ADVP
VP
.
NP
VP
PU
S

Basic Node Alignment



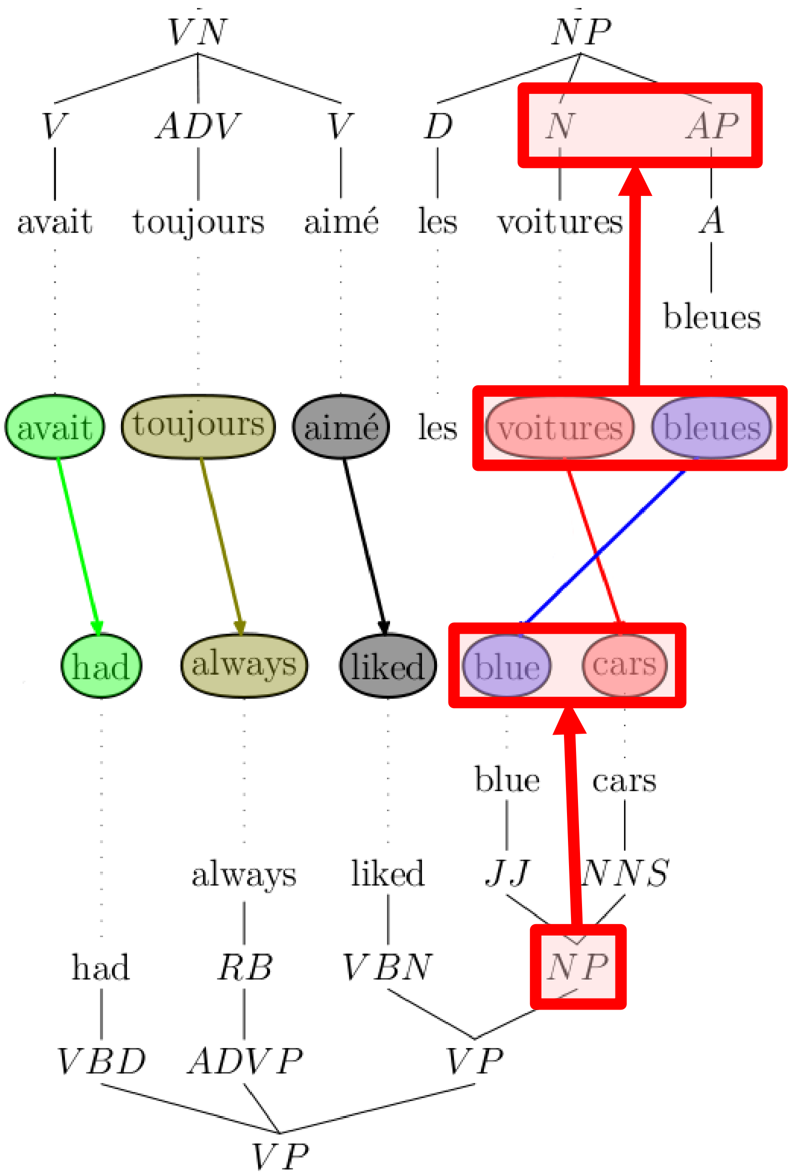
- Word alignment consistency constraint from phrase-based SMT

Basic Node Alignment



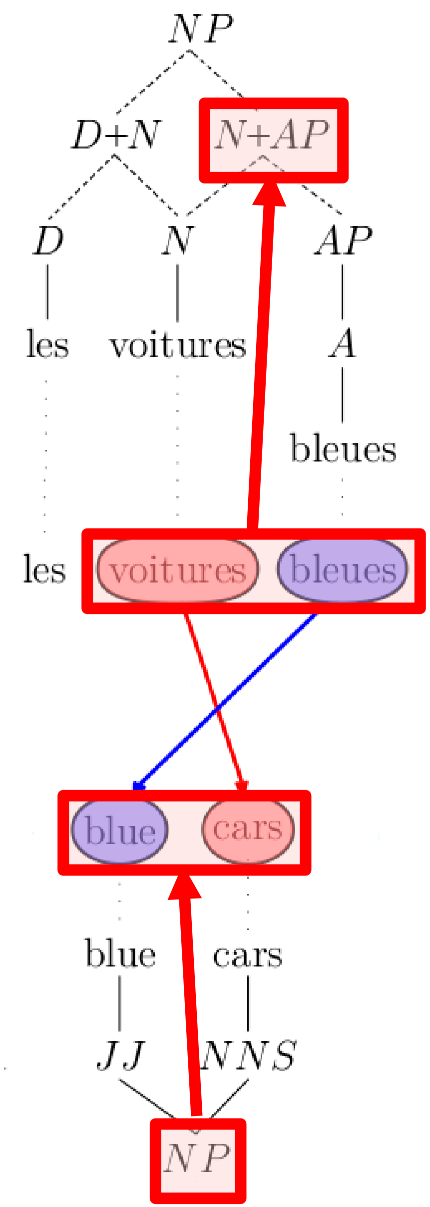
- Word alignment consistency constraint from phrase-based SMT

Virtual Nodes



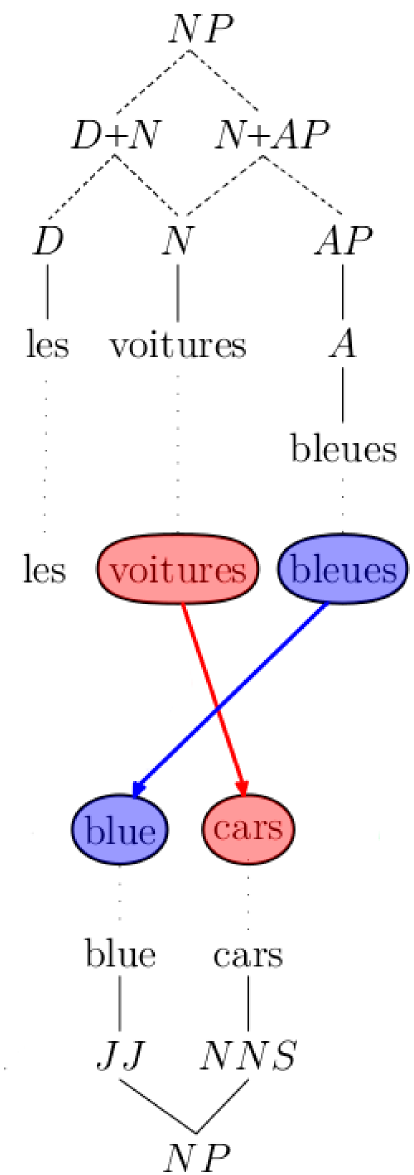
- Consistently aligned consecutive children of the same parent

Virtual Nodes



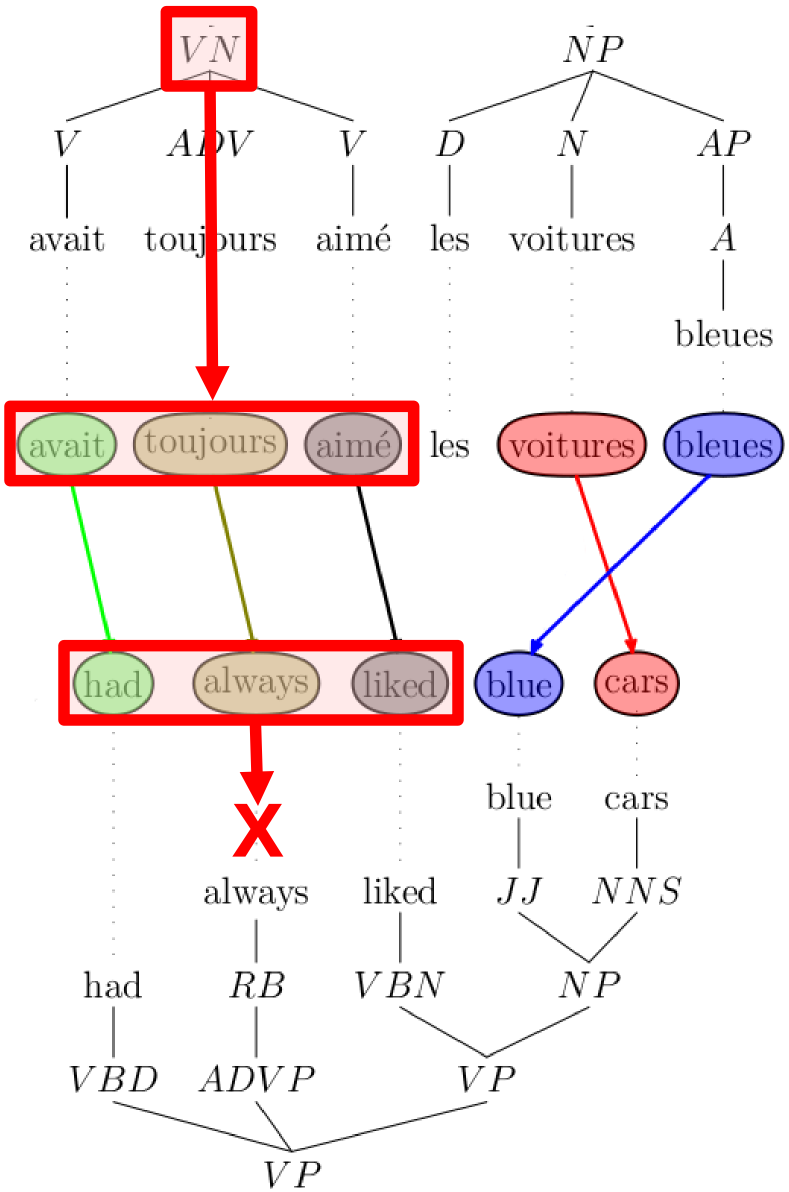
- Consistently aligned consecutive children of the same parent
- New intermediate node inserted in tree

Virtual Nodes



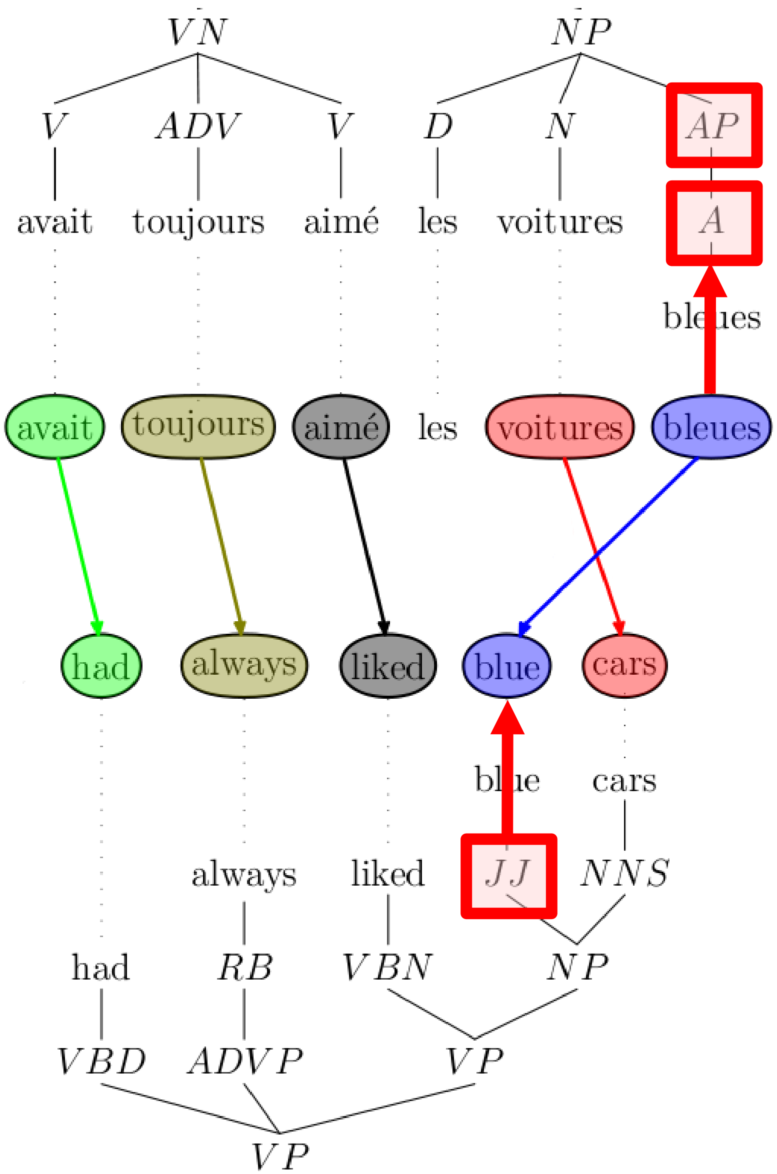
- Consistently aligned consecutive children of the same parent
- New intermediate node inserted in tree
- Virtual nodes may overlap
- Virtual nodes may align to any type of node

Syntax Constraints



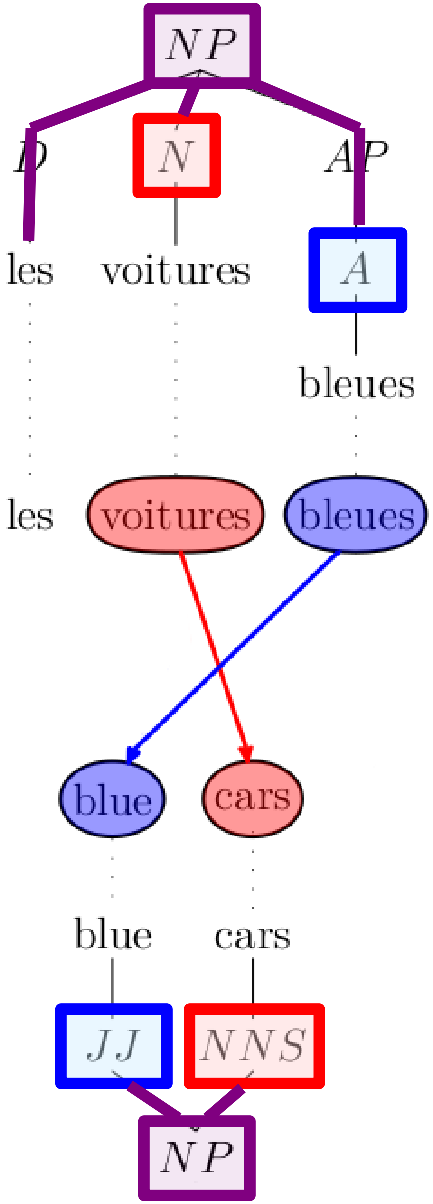
- Consistent word alignments ≠ node alignment
- Virtual nodes may not cross constituent boundaries

Multiple Alignment



- Nodes with multiple consistent alignments
- Keep all of them

Basic Grammar Extraction



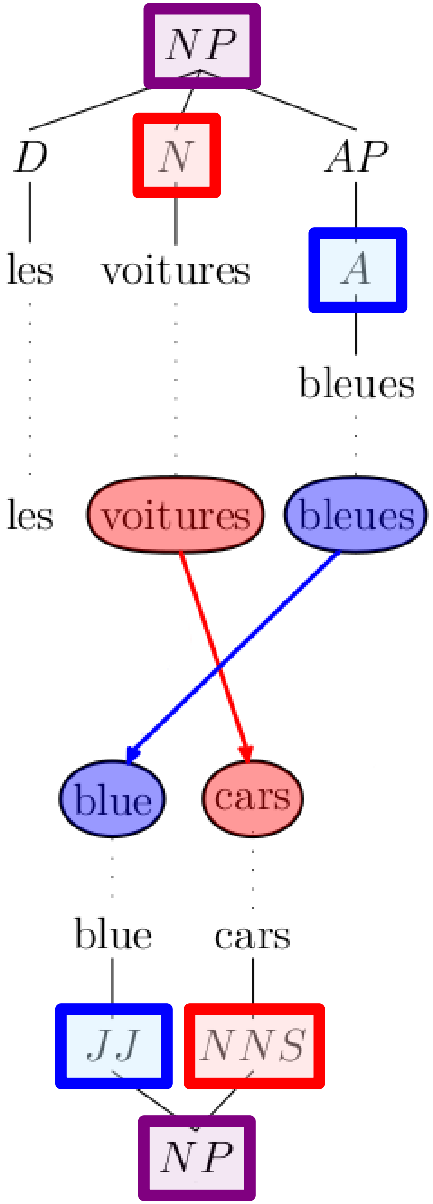
- Aligned node pair is LHS; aligned subnodes are RHS

$$\mathrm{NP::NP} \rightarrow [\mathrm{les\ N^1\ A^2}]::[\mathrm{JJ^2\ NNS^1}]$$

$$\mathrm{N::NNS} \rightarrow [\mathrm{voitures}]::[\mathrm{cars}]$$

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

Multiple Decompositions



- All possible right-hand sides are extracted

NP::NP → [les N^1 A^2]::[JJ^2 NNS^1]

NP::NP → [les N^1 bleues]::[blue NNS^1]

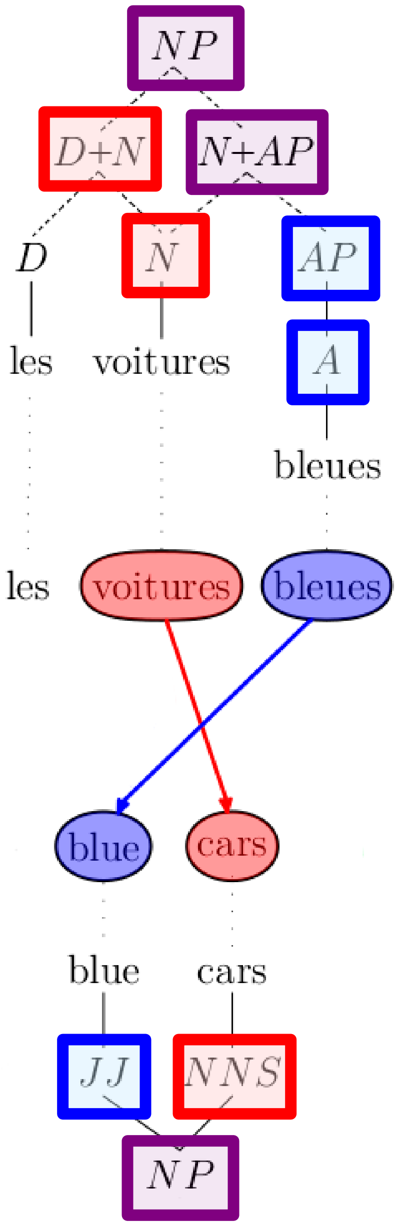
NP::NP → [les voitures A^2]::[JJ^2 cars]

NP::NP → [les voitures bleues]::[blue cars]

N::NNS → [voitures]::[cars]

A::JJ → [bleues]::[blue]

Multiple Decompositions



NP::NP $\rightarrow$ [les N+AP^{1}]::[NP^{1}]

NP::NP $\rightarrow$ [D+N^{1} AP^{2}]::[JJ^{2} NNS^{1}]

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

NP::NP $\rightarrow$ [les N^{1} AP^{2}]::[JJ^{2} NNS^{1}]

NP::NP $\rightarrow$ [les N^{1} A^{2}]::[JJ^{2} NNS^{1}]

NP::NP $\rightarrow$ [D+N^{1} bleues]::[blue NNS^{1}]

NP::NP $\rightarrow$ [les N^{1} bleues]::[blue NNS^{1}]

NP::NP $\rightarrow$ [les voitures AP^{2}]::[JJ^{2} cars]

NP::NP $\rightarrow$ [les voitures A^{2}]::[JJ^{2} cars]

NP::NP $\rightarrow$ [les voitures bleues]::[blue cars]

D+N::NNS $\rightarrow$ [les N^{1}]::[NNS^{1}]

D+N::NNS $\rightarrow$ [les voitures]::[cars]

N+AP::NP $\rightarrow$ [N^{1} AP^{2}]::[JJ^{2} NNS^{1}]

N+AP::NP $\rightarrow$ [N^{1} A^{2}]::[JJ^{2} NNS^{1}]

N+AP::NP $\rightarrow$ [N^{1} bleues]::[blue NNS^{1}]

N+AP::NP $\rightarrow$ [voitures AP^{2}]::[JJ^{2} cars]

N+AP::NP $\rightarrow$ [voitures A^{2}]::[JJ^{2} cars]

N+AP::NP $\rightarrow$ [voitures bleues]::[blue cars]

N::NNS $\rightarrow$ [voitures]::[cars]

AP::JJ $\rightarrow$ [A^{1}]::[JJ^{1}]

AP::JJ $\rightarrow$ [bleues]::[blue]

A::JJ $\rightarrow$ [bleues]::[blue]

Constraints

- Max rank of phrase pair rules
- Max rank of hierarchical rules
- Max number of siblings in a virtual node
- Whether to allow unary chain rules

$$\mathrm{NP::NP} \rightarrow [\mathrm{PRO}^1]::[\mathrm{PRP}^1]$$

- Whether to allow “triangle” rules

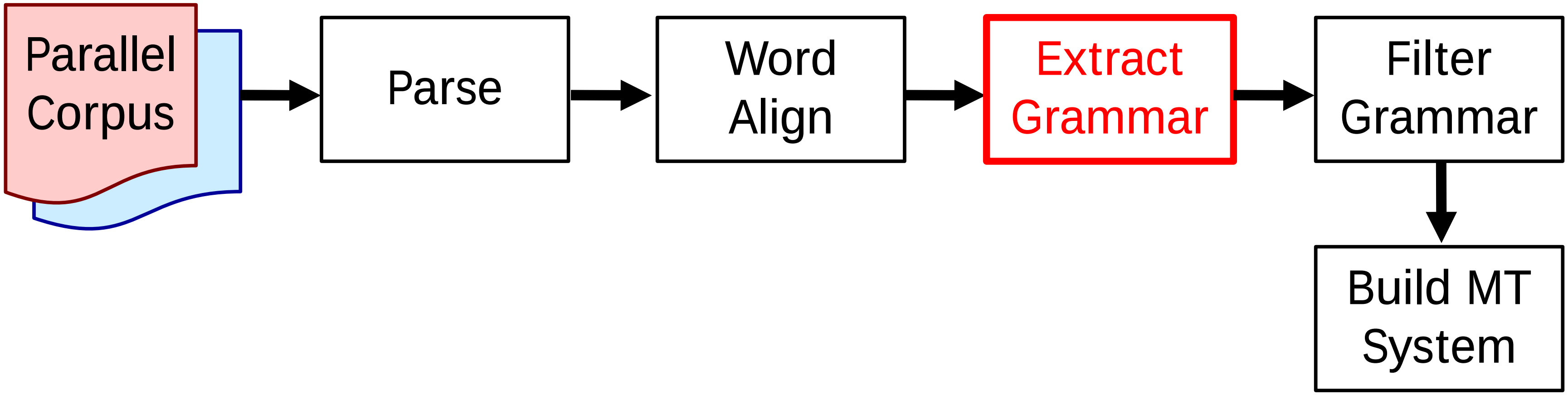
$$\mathrm{AP::JJ} \rightarrow [\mathrm{A}^1]::[\mathrm{JJ}^1]$$

Comparison to Related Work

	Tree Constr.	Multiple Aligns	Virtual Nodes	Multiple Decomp.
Hiero	No	—	—	Yes
Stat-XFER	Yes	No	Some	No
GHKM	Yes	No	No	Yes
SAMT	No	No	Yes	Yes
Chiang [2010]	No	No	Yes	Yes
This work	Yes	Yes	Yes	Yes

Experimental Setup

- Train: FBIS Chinese–English corpus
- Tune: NIST MT 2006
- Test: NIST MT 2003



Extraction Configurations

- Baseline:
  - Stat-XFER exact tree-to-tree extractor
  - Single decomposition with minimal rules
- Multi:
  - Add multiple alignments and decompositions
- Virt short:
  - Add virtual nodes; max rule length 5
- Virt long:
  - Max rule length 7

Number of Rules Extracted

	Tokens		Types	
	Phrase	Hierarc.	Phrase	Hierarc.
Baseline	6,646,791	1,876,384	1,929,641	767,573
Multi	8,709,589	6,657,590	2,016,227	3,590,184
Virt short	10,190,487	14,190,066	2,877,650	8,313,690
Virt long	10,288,731	22,479,863	2,970,403	15,750,695

Number of Rules Extracted

	Tokens		Types	
	Phrase	Hierarc.	Phrase	Hierarc.
Baseline	6,646,791	1,876,384	1,929,641	767,573
Multi	8,709,589	6,657,590	2,016,227	3,590,184
Virt short	10,190,487	14,190,066	2,877,650	8,313,690
Virt long	10,288,731	22,479,863	2,970,403	15,750,695

- Multiple alignments and decompositions:
  - Four times as many hierarchical rules
  - Small increase in number of phrase pairs

Number of Rules Extracted

	Tokens		Types	
	Phrase	Hierarc.	Phrase	Hierarc.
Baseline	6,646,791	1,876,384	1,929,641	767,573
Multi	8,709,589	6,657,590	2,016,227	3,590,184
Virt short	10,190,487	14,190,066	2,877,650	8,313,690
Virt long	10,288,731	22,479,863	2,970,403	15,750,695

- Multiple decompositions and virtual nodes:
  - 20 times as many hierarchical rules
  - Stronger effect on phrase pairs
  - 46% of rule types use virtual nodes

Number of Rules Extracted

	Tokens		Types	
	Phrase	Hierarc.	Phrase	Hierarc.
Baseline	6,646,791	1,876,384	1,929,641	767,573
Multi	8,709,589	6,657,590	2,016,227	3,590,184
Virt short	10,190,487	14,190,066	2,877,650	8,313,690
Virt long	10,288,731	22,479,863	2,970,403	15,750,695

- Proportion of singletons mostly unchanged
- Average hierarchical rule count drops

Rule Filtering for Decoding

- All phrase pair rules that match test set
- Most frequent hierarchical rules:
  - Top 10,000 of all types
  - Top 100,000 of all types
  - Top 5,000 fully abstract
    + top 100,000 partially lexicalized

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

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

Results: Metric Scores

- NIST MT 2003 test set

System	Filter	BLEU	METR	TER
Baseline	**10k**	24.39	54.35	68.01
Multi	**10k**	24.28	53.58	65.30
Virt short	**10k**	25.16	54.33	66.25
Virt long	**10k**	25.74	54.55	65.52

- Strict grammar filtering: extra phrase pairs help improve scores

Results: Metric Scores

- NIST MT 2003 test set

System	Filter	BLEU	METR	TER
Baseline	**5k+100k**	25.95	54.77	66.27
Virt short	**5k+100k**	26.08	54.58	64.32
Virt long	**5k+100k**	25.83	54.35	64.55

- Larger grammars: score difference erased

Conclusions

- Very large linguistically motivated rule sets
  - No violating constituent bounds (Stat-XFER)
  - Multiple node alignments
  - Multiple decompositions (Hiero, GHKM)
  - Virtual nodes (< SAMT)
- More phrase pairs help improve scores
- Grammar filtering has significant impact

Automatic Category Label Coarsening for Syntax-Based Machine Translation

Joint work with Greg Hanneman

Motivation

- S-CFG-based MT:
  - Training data annotated with constituency parse trees on both sides
  - Extract labeled S-CFG 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 sub-optimal

Motivation

- Category labels have significant impact on syntax-based MT
  - Govern which rules can combine together
  - Generate derivational ambiguity
  - Fragment the data during rule acquisition
  - Greatly impact decoding complexity
- Granularity spectrum has ranged from single category (Chiang's Hiero) to 1000s of labels (SAMT, our new Rule Learner)
- Our default category labels are artifacts of the underlying monolingual parsers used
  - Based on TreeBanks, designed independently for each language, without MT in mind
  - Not optimal even for monolingual parsing
  - What labels are necessary and sufficient for effective syntax-based decoding?

Research Goals

- Define and measure the effect labels have
  - Spurious ambiguity, rule sparsity, and reordering precision
- Explore the space of labeling schemes
  - Collapsing labels

    NN
    NNS $\rightarrow$ N

  - Refining labels

    JJ::A $\rightarrow$ JJ::AA
    JJ::AB

  - Correcting local labeling errors

    PRO $\rightarrow$ N

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]

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

$$VP::VP \rightarrow [VV^1\ VV^2]::[RB^1\ VBN^2]$$

Motivation

- **Rule sparsity:**
  - Label mismatch blocks rule application

$$VP::VP \rightarrow [VV^1 \text{了} PP^2 \text{的} NN^3]::[VBD^1 \text{ their } NN^3 \; PP^2]$$

$$VP::VP \rightarrow [VV^1 \text{了} PP^2 \text{的} NN^3]::[VB^1 \text{ their } NNS^3 \; PP^2]$$

+ 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

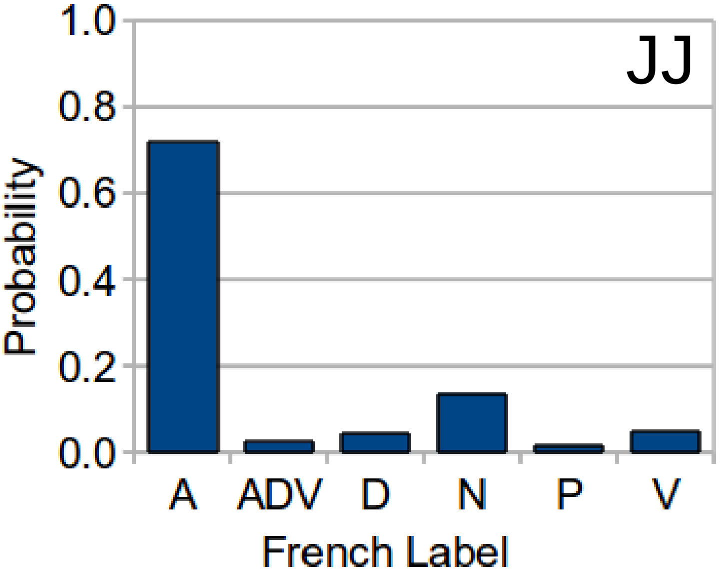
la grande voiture	the large car
la plus grande voiture	the larger car
la voiture la plus grande	the largest car

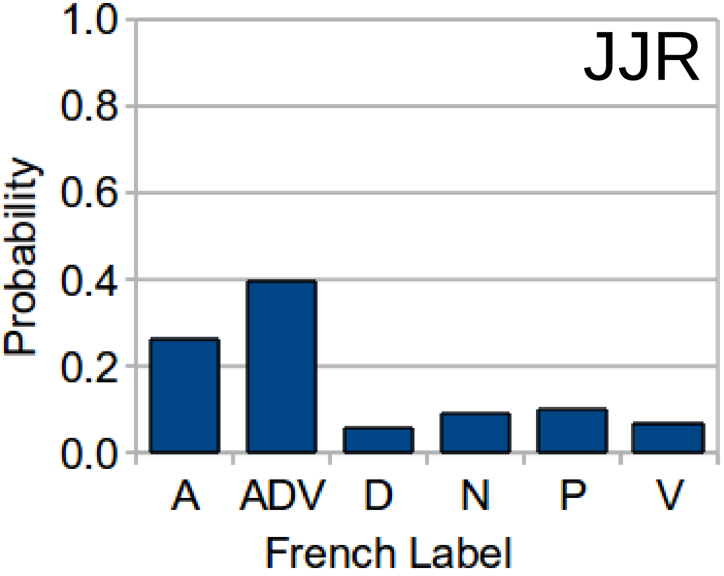
- Simple measure: how category is aligned to other language

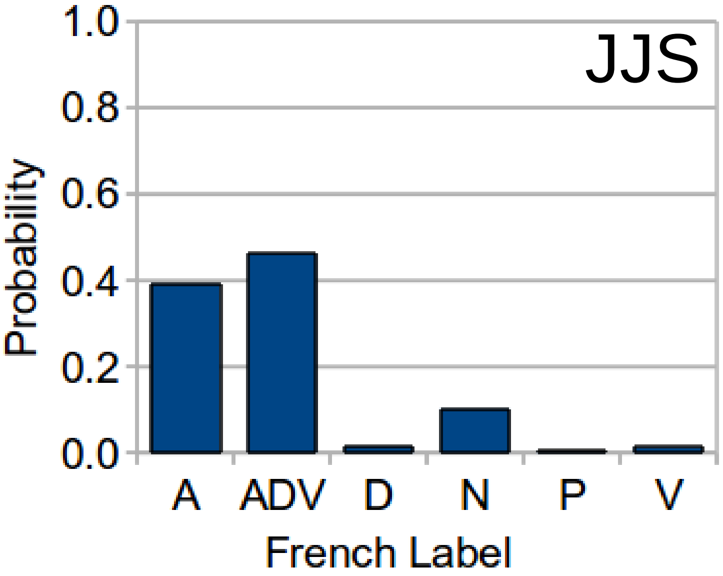
A::JJ → [grande]::[large]

AP::JJR → [plus grande]::[larger]

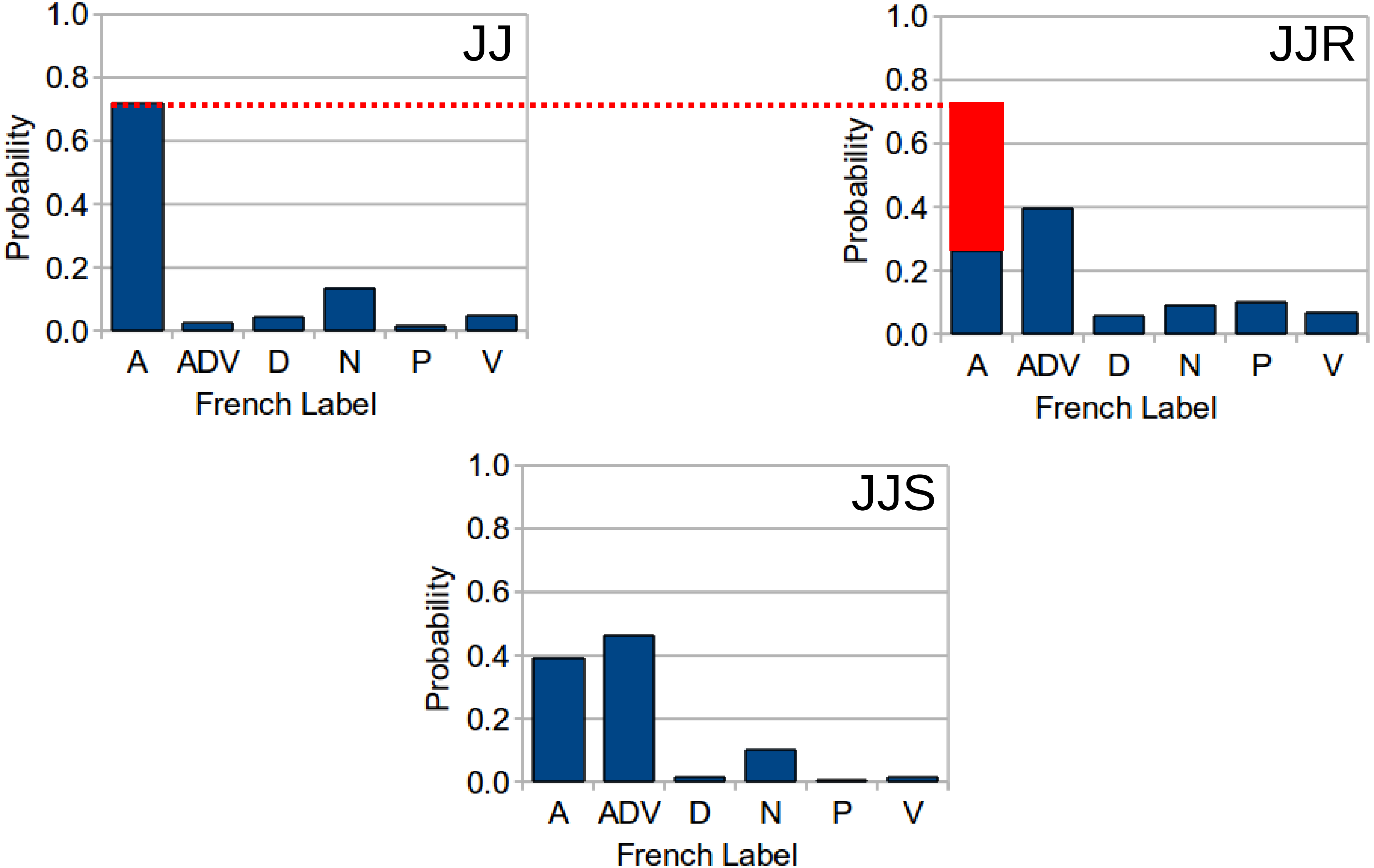
L_1 Alignment Distance



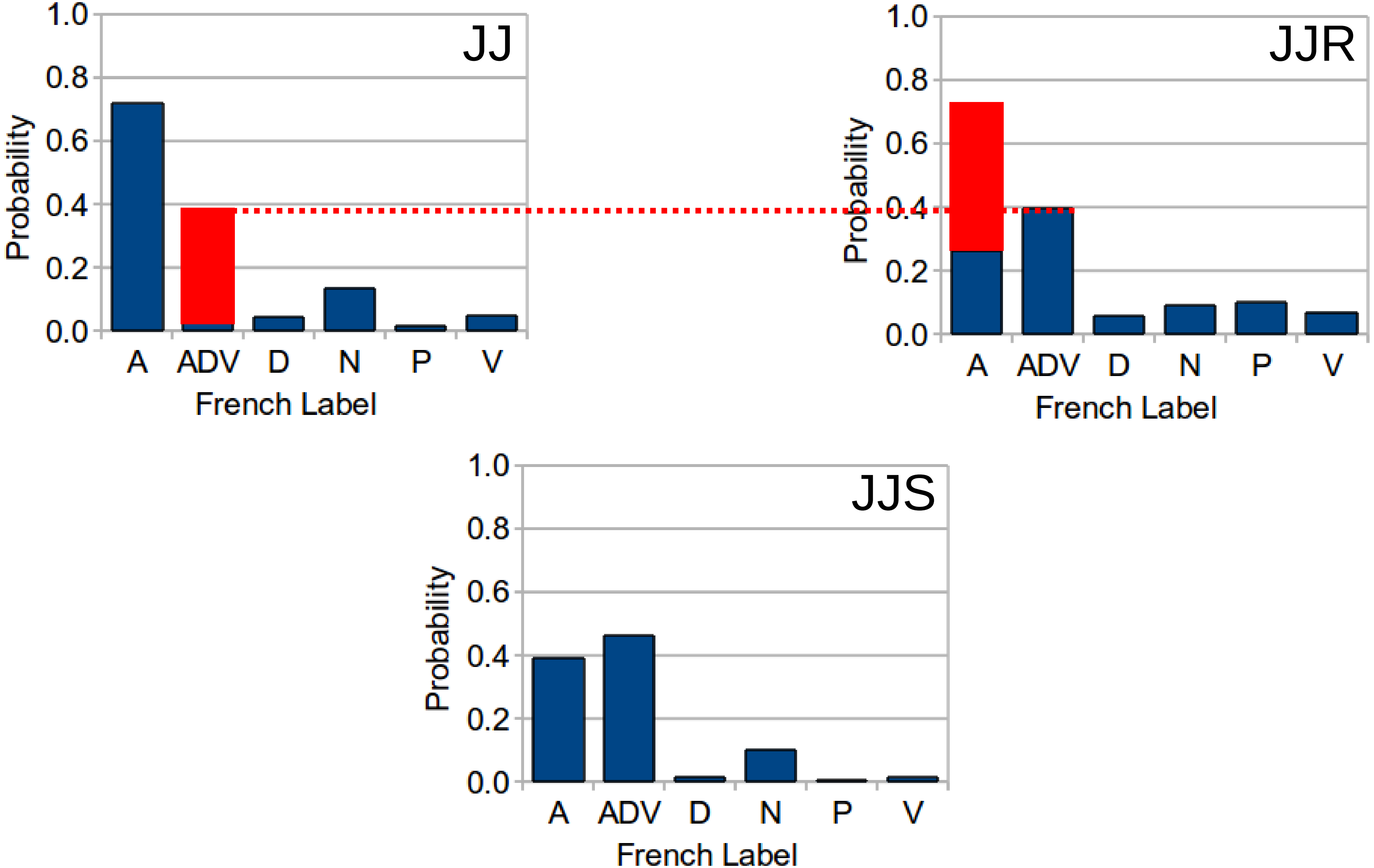




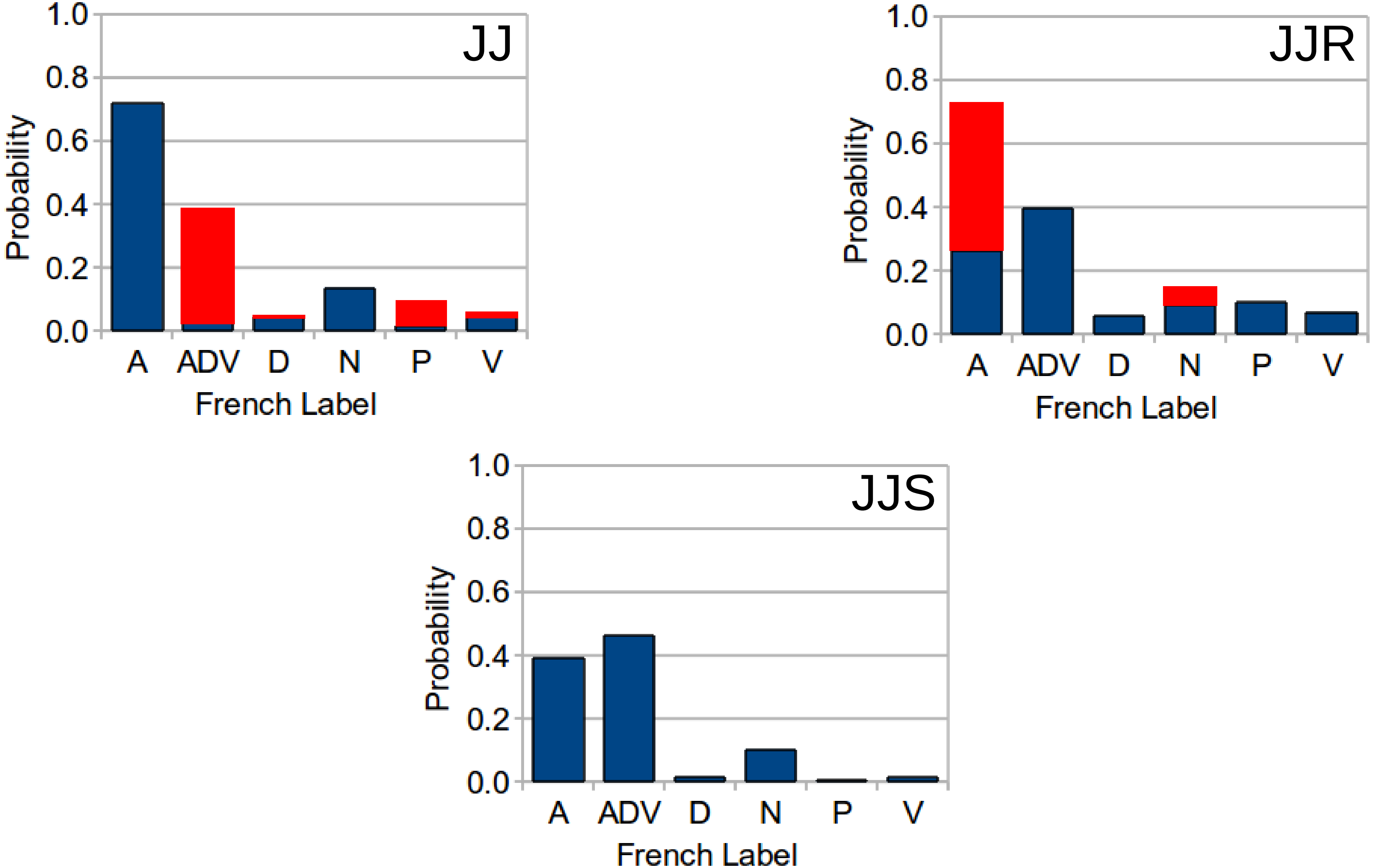
L_1 Alignment Distance



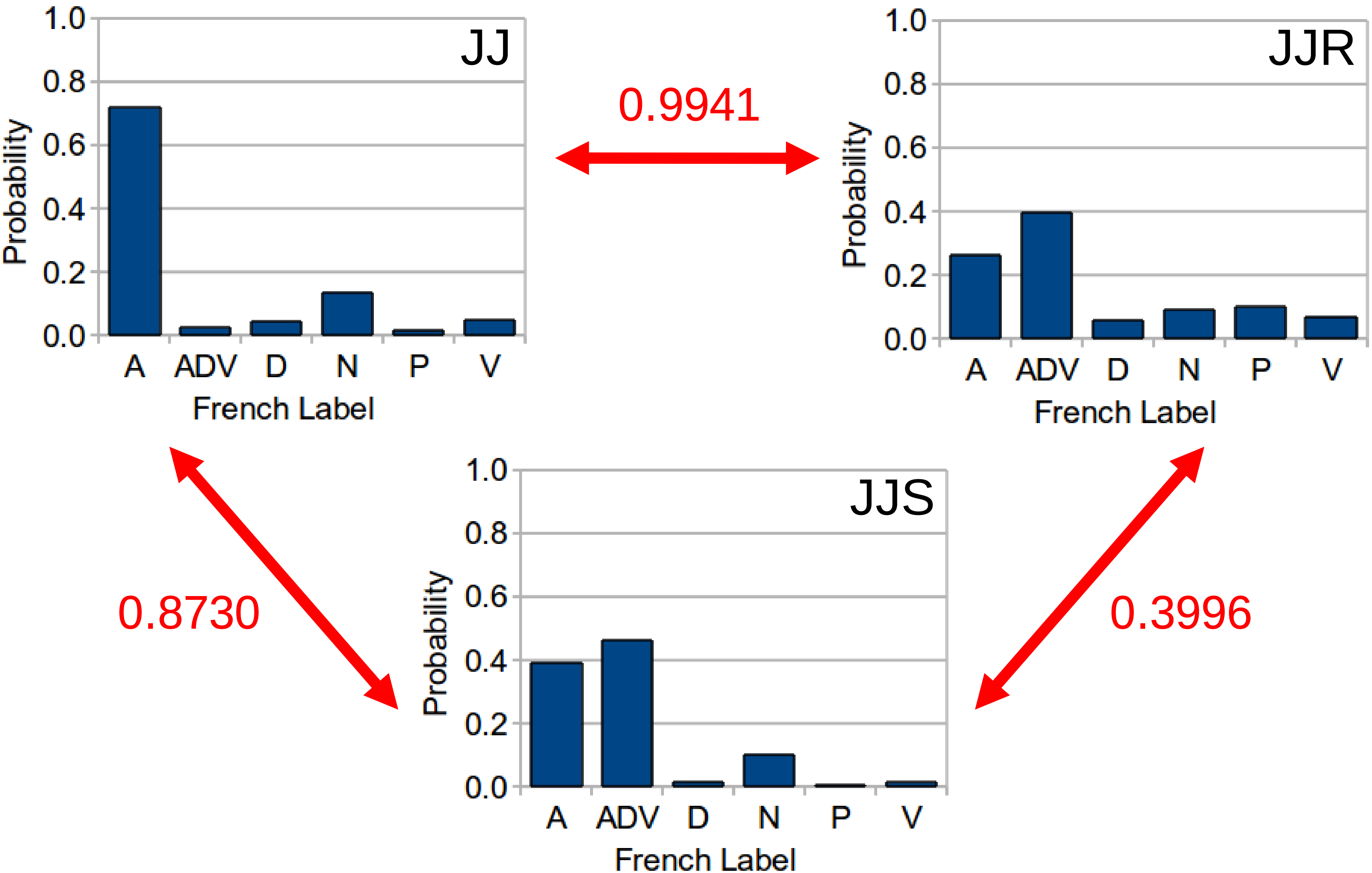
L_1 Alignment Distance



L_1 Alignment Distance



L_1 Alignment Distance

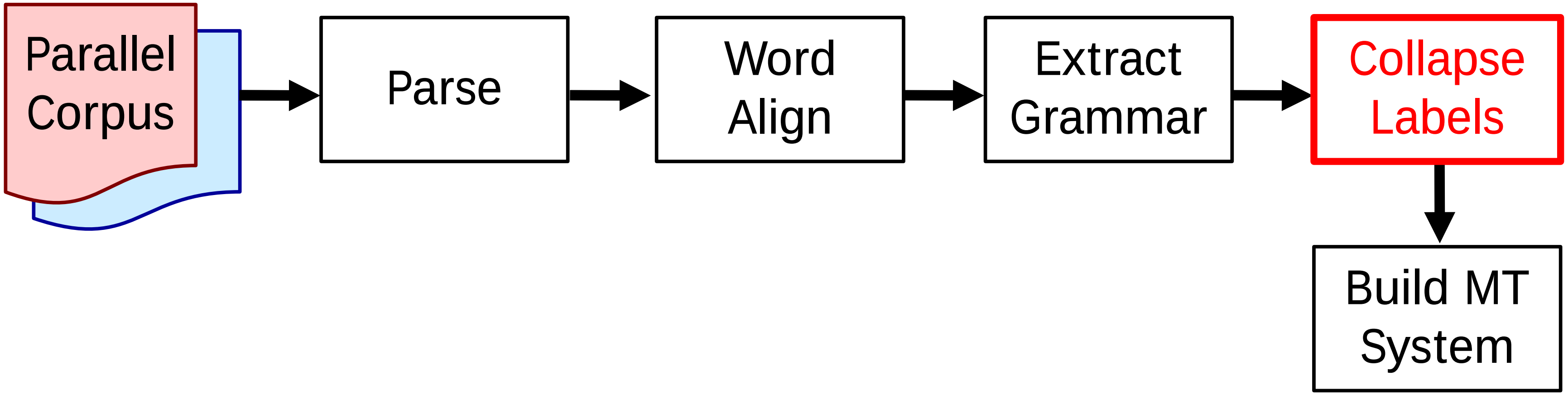


Label Collapsing Algorithm

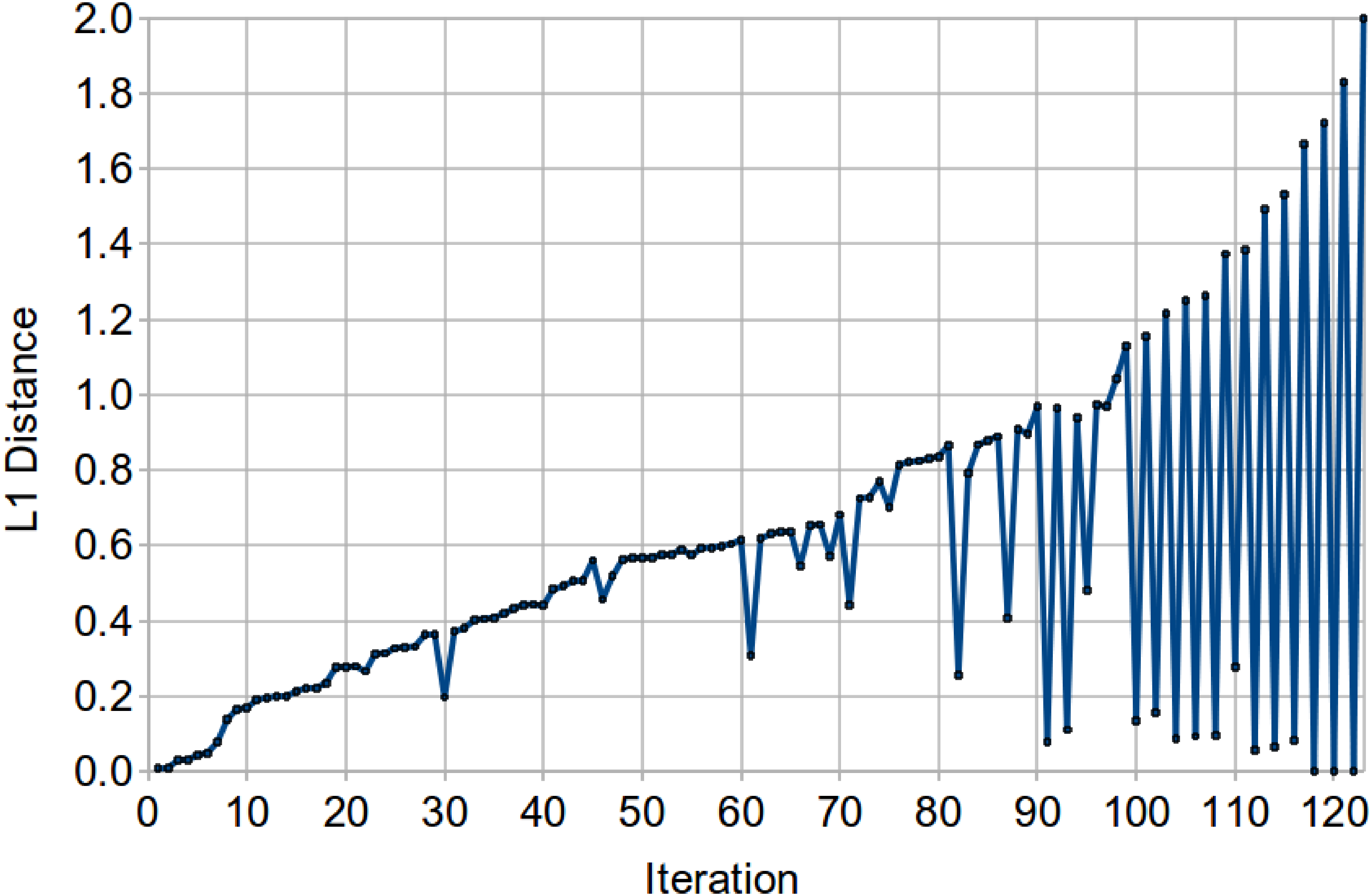
- Extract baseline grammar from aligned tree pairs (e.g. Lavie et al. [2008])
- Compute label alignment distributions
- Repeat until stopping point:
  - Compute L_1 distance between all pairs of source and target labels
  - Merge the label pair with smallest distance
  - Update label alignment distributions

Experiment 1

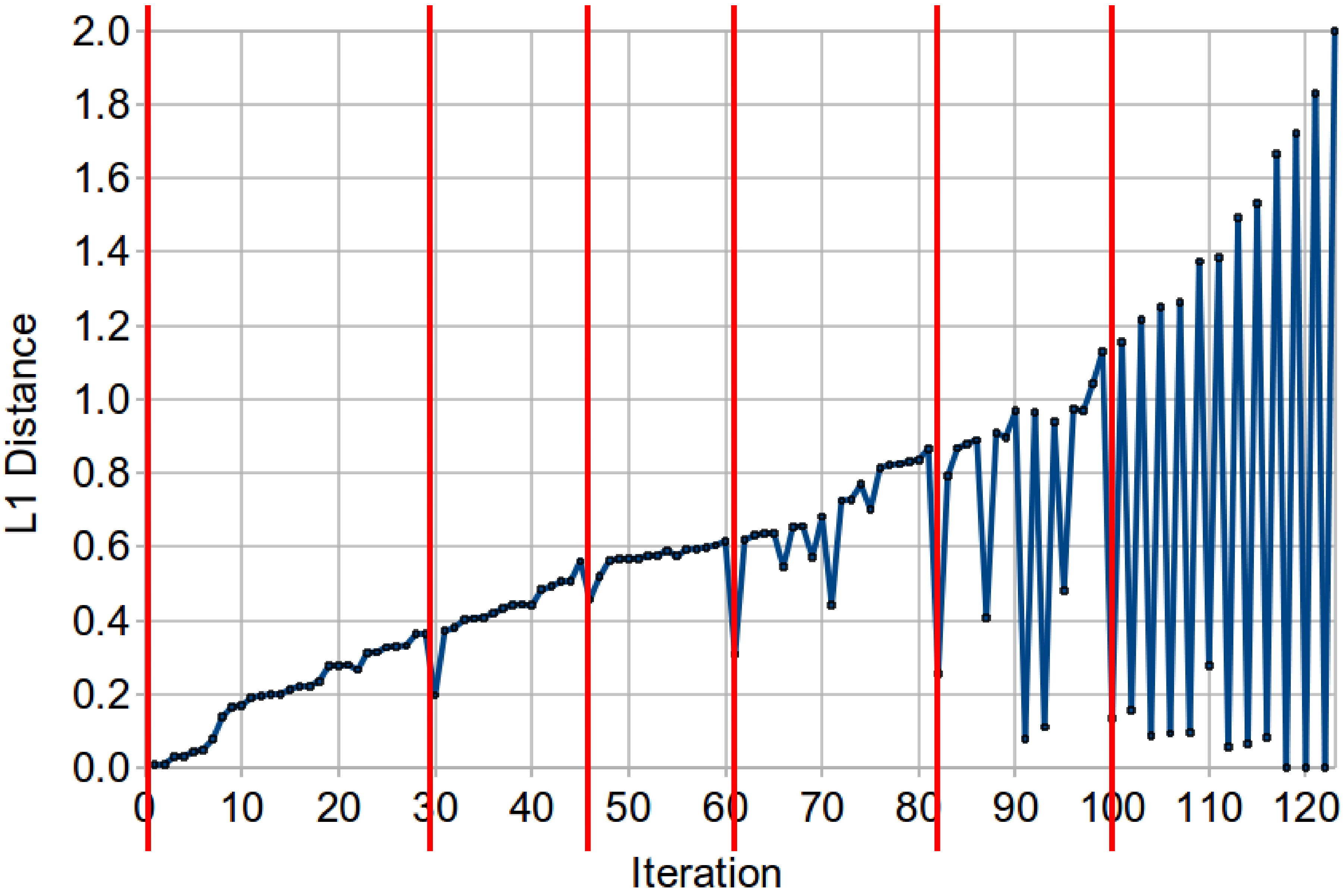
- Goal: Explore effect of collapsing with respect to stopping point
- Data: Chinese–English FBIS corpus (302 k)



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

    NN::NN → [友好]::[friendship]

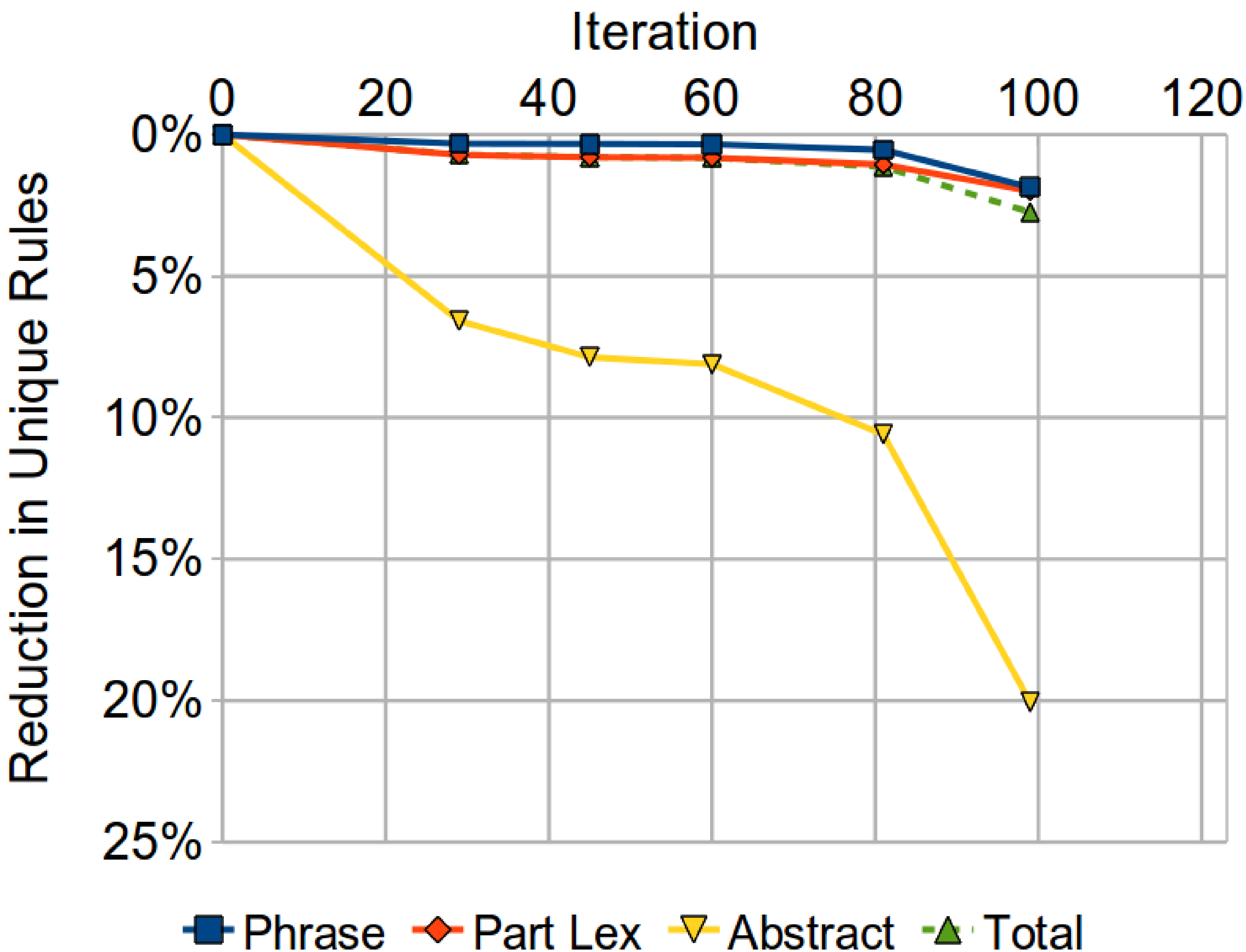
  - Partially lexicalized grammar rules

    NP::NP → [2000年 NN^1]::[the 2000 NN^1]

  - Fully abstract grammar rules

    VP::ADJP → [VV^1 VV^2]::[RB^1 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

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

Summary and Conclusions

- Increasing consensus in the MT community on the necessity of models that integrate deeper-levels of linguistic analysis and abstraction
  - Especially for languages with rich morphology and for language pairs with highly-divergent syntax
- Progress has admittedly been slow
  - No broad understanding yet of what we should be modeling and how to effectively acquire it from data
  - Challenges in accurate annotation of vast volumes of parallel training data with morphology and syntax
  - What is necessary and effective for monolingual NLP isn't optimal or effective for MT
  - Complexity of Decoding with these types of models
- Some insights and (partial) solutions
- Lots of interesting research forthcoming

References

- Al-Haj, H. and A. Lavie. "The Impact of Arabic Morphological Segmentation on Broad-coverage English-to-Arabic Statistical Machine Translation". In Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA-2010), Denver, Colorado, November 2010.
- Al-Haj, H. and A. Lavie. "The Impact of Arabic Morphological Segmentation on Broad-coverage English-to-Arabic Statistical Machine Translation“. MT Journal Special Issue on Arabic MT. Under review.

References

- Chiang (2005), "A hierarchical phrase-based model for statistical machine translation," ACL
- Chiang (2010), "Learning to translate with source and target syntax," ACL
- Galley, Hopkins, Knight, and Marcu (2004), "What's in a translation rule?," NAACL
- Lavie, Parlikar, and Ambati (2008), "Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora," SSST-2
- Zollmann and Venugopal (2006), "Syntax augmented machine translation via chart parsing," WMT

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