Phrase-Based Statistical Machine Translation with Pivot Languages

N. Bertoldi, †M. Barbaiani, M. Federico, R. Cattoni

FBK, Trento - Italy
† Rovira i Virgili University, Tarragona - Spain

October 21st, 2008

Pivot Translation

- **Assumptions**:
  - no parallel data between source language $\mathcal{F}$ and target language $\mathcal{E}$
  - two independent parallel corpora (F,G_F) and (G_E,E)
  - two full-fledged MT systems $\mathcal{F} \rightarrow \mathcal{G}$ and $\mathcal{G} \rightarrow \mathcal{E}$
- **Problem**: how to perform translation from $\mathcal{F}$ to $\mathcal{E}$?
- **Approach 1:** Bridging at translation time

source text	$\mathcal{F} \rightarrow \mathcal{G}$	pivot text	$\mathcal{G} \rightarrow \mathcal{E}$	target text
$\mathbf{f}$	$\rightarrow$	$\mathbf{g}$	$\rightarrow$	$\mathbf{e}$

- **Approach 2:** Bridging at training time

synthetic training data	generated by translating	with system
(F,$\bar{E}_F$)	G_F of (F,G_F)	$\mathcal{G} \rightarrow \mathcal{E}$
($\bar{F}_E$,E)	G_E of (G_E,E)	$\mathcal{G} \rightarrow \mathcal{F}$

Pivot Task description

- BTEC domain data
- Pivot Task of IWSLT 2008: Chinese-English-Spanish

- training data: CE1, CE2, ES1, and CS1 (19K sentences)
- disjoint condition: CE2 and ES1
- overlap condition: CE1 and ES1
- direct condition: CS1

- dev set: 506 Chinese sentences with 7 refs in English and Spanish
- test set: 1K sentences with 1 reference extracted from CES1

Statistical Machine Translation

$$\begin{array}{ccc} \text{source text} & & \text{target text} \\ \mathbf{f} & \longrightarrow & \mathbf{e} \end{array}$$

- alignment-based parametric model

$$p(\mathbf{e} \mid \mathbf{f}) = \sum_{\mathbf{a}} p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) = \sum_{\mathbf{a}} p_{\theta_{FE}}(\mathbf{e}, \mathbf{a} \mid \mathbf{f})$$

- parameter estimation:

$$\hat{\theta}_{FE} = \arg \underset{\theta_{FE}}{max} \prod_i p_{\theta_{FE}}(\mathbf{e}_i \mid \mathbf{f}_i) \qquad \text{given } \{(\mathbf{f}_i, \mathbf{e}_i)\}$$

- search criterion:

$$\mathbf{f} \rightarrow \hat{\mathbf{e}} \approx \arg \underset{\mathbf{e}}{max} \, \underset{\mathbf{a}}{max} \, p_{\hat{\theta}_{FE}}(\mathbf{e}, \mathbf{a} \mid \mathbf{f})$$

Direct baseline system

- open-source MT toolkit **Moses**
- statistical **log-linear** model with 8 features
- weight optimization by means of a **minimum error training** procedure

- **phrase-based** translation model:
  - direct and inverted frequency-based and lexical-based probabilities
  - phrase pairs extracted from symmetrized word alignments (GIZA++)
- 5-gram word-based LM exploiting Improved Kneser-Ney smoothing (IRSTLM)
- standard negative-exponential distortion model
- word and phrase penalties

Bridging at translation time

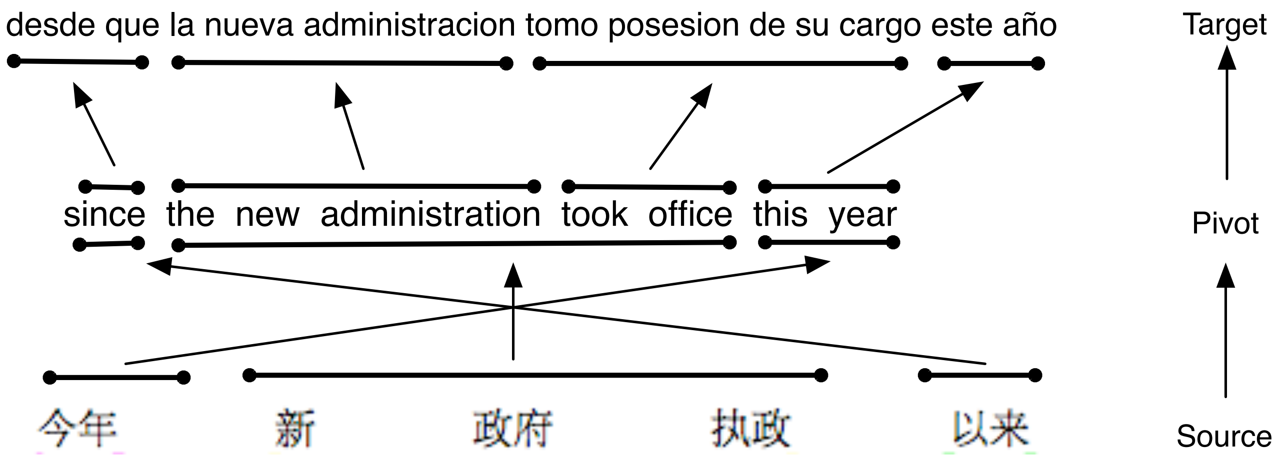
$$\begin{array}{ccccc} \text{source text} & & \text{pivot text} & & \text{target text} \\ \mathbf{f} & \rightarrow & \mathbf{g} & \rightarrow & \mathbf{e} \end{array}$$

$$\begin{aligned} p(\mathbf{e} \mid \mathbf{f}) &= \sum_{\mathbf{g}} p(\mathbf{e}, \mathbf{g} \mid \mathbf{f}) = \sum_{\mathbf{g}} p(\mathbf{g} \mid \mathbf{f})\ p(\mathbf{e} \mid \mathbf{g}) \\ &= \sum_{\mathbf{g}} \sum_{\mathbf{b}} p_{\theta_{FG}}(\mathbf{g}, \mathbf{b} \mid \mathbf{f}) \ \sum_{\mathbf{a}} p_{\theta_{GE}}(\mathbf{e}, \mathbf{a} \mid \mathbf{g}) \end{aligned}$$

$$\mathbf{f} \rightarrow \hat{\mathbf{e}} \quad \approx \quad \arg \underset{\mathbf{e},\mathbf{g}}{max}\ \underset{\mathbf{a},\mathbf{b}}{max}\ \ p_{\hat{\theta}_{FG}}(\mathbf{g}, \mathbf{b} \mid \mathbf{f}) p_{\hat{\theta}_{GE}}(\mathbf{e}, \mathbf{a} \mid \mathbf{g})$$

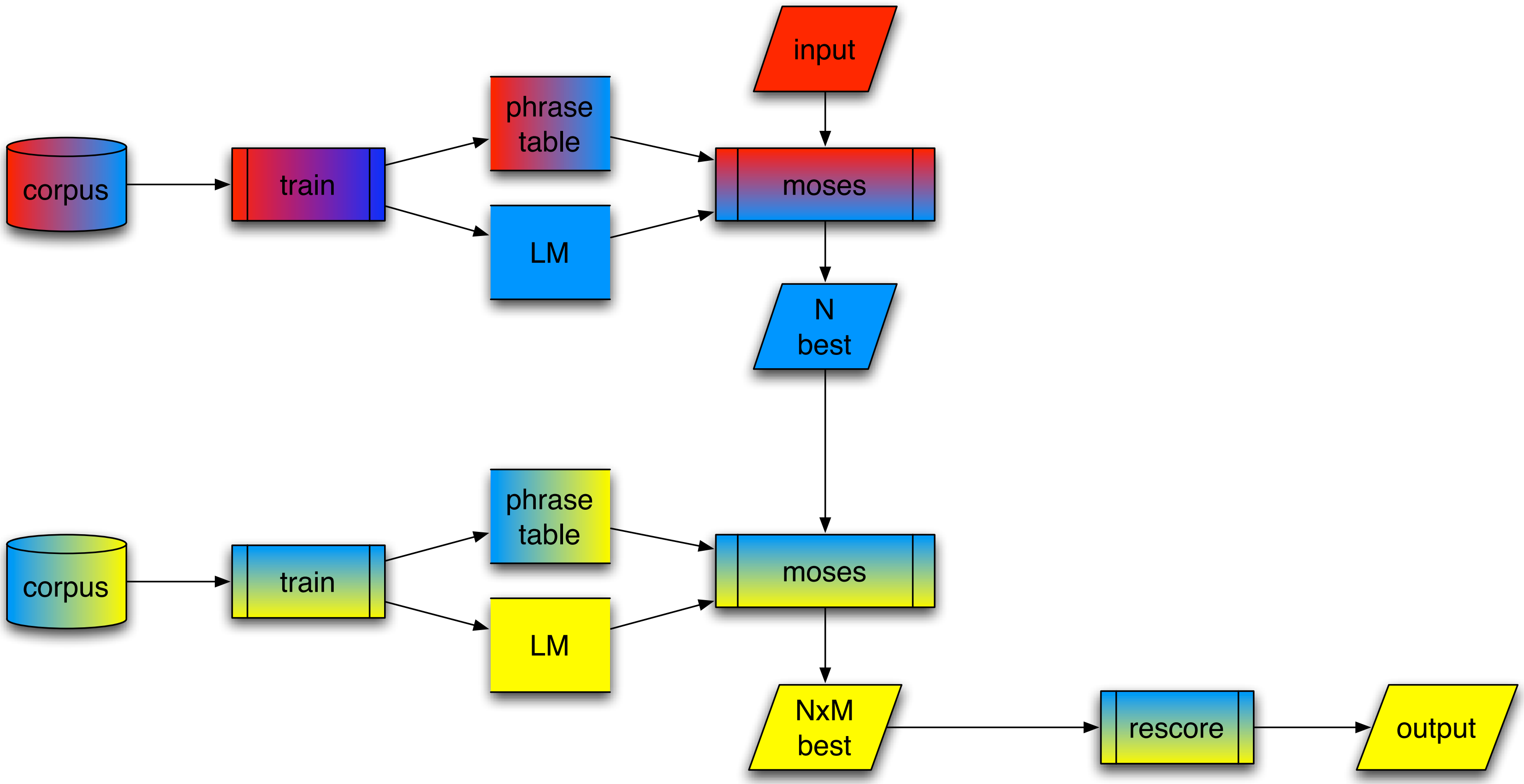
- two full-fedged systems trained on corpora (F,G$_F$) and (G$_E$,E)
- search including the pivot language increases complexity

Coupling with Unconstrained Alignments



- **sentence-level** coupling
- requires performing search over two alignments
- search can be decoupled over a subset of hypotheses:
  - N-best list (or word lattices) of source-to-pivot translations

Coupling with Unconstrained Alignments



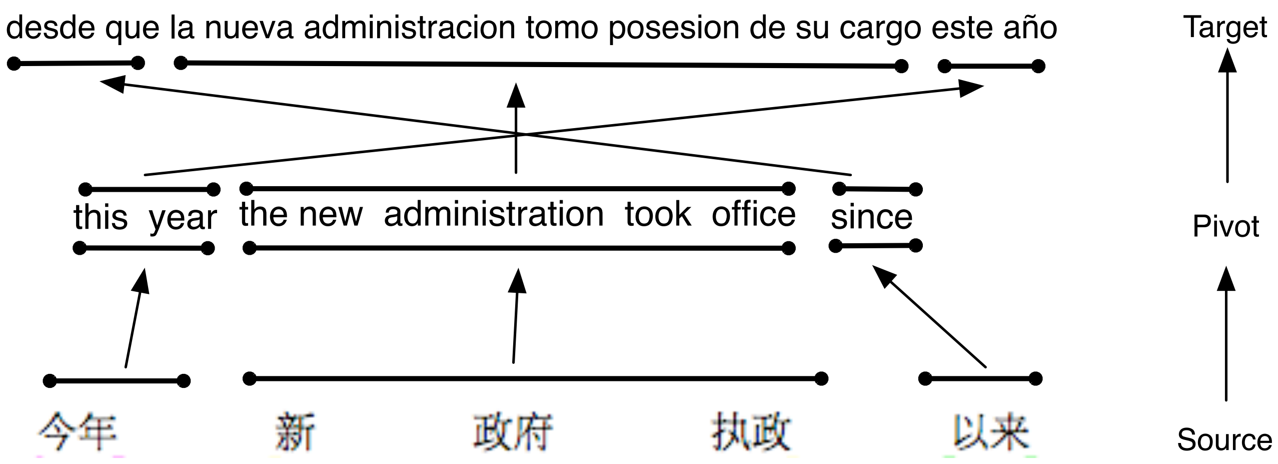
Coupling with Unconstrained Alignments

n,m	rescoring features	dev	test
1	-	25.13	16.44
10	2	25.28	16.60
	16	26.65	17.59
20	16	27.18	17.03
50	16	27.78	16.96
100	16	27.89	17.64

- 16 feature scores $>$ 2 global scores
- 100x100-best gives best performance on dev set

- time expensive: $(N+1)$ translation + rescoring

Coupling with Constrained Alignments



- **phrase-level** coupling
- share segmentation on the pivot language and use just one re-ordering
- needs one distortion model that directly models source to target
- needs only one target language model

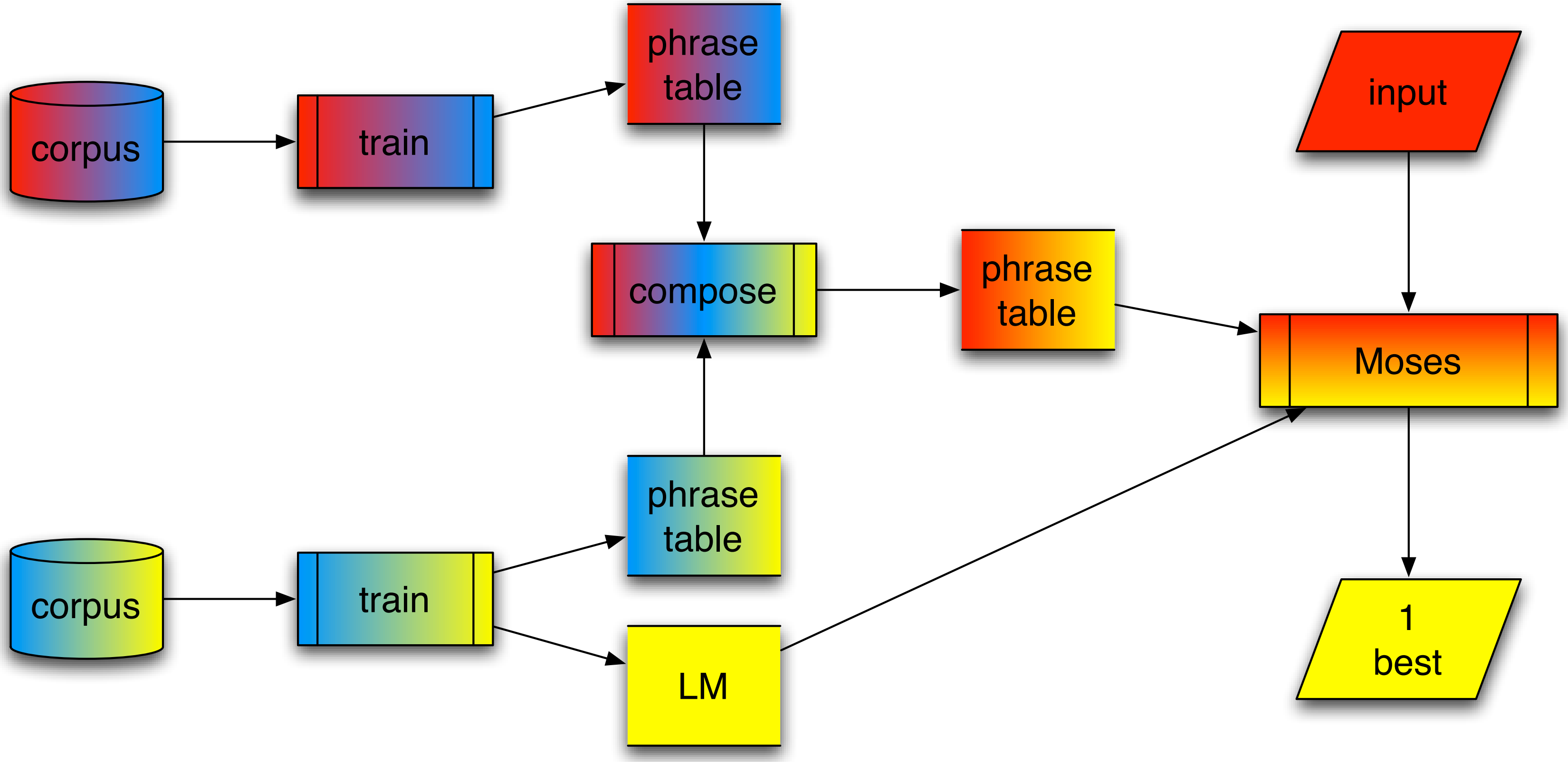
Coupling with Constrained Alignments

- needs to modify decoder, or
- compose phrase table before decoding

$$\begin{aligned} PT(F,E) &= PT(F,G) \otimes PT(G,E) \\ &= \{(\tilde{f},\tilde{e}) \mid \exists\ \tilde{g} \text{ s.t. } (\tilde{f},\tilde{g}) \in PT(F,G_F) \wedge \exists\ (\tilde{g},\tilde{e}) \in PT(G_E,E)\end{aligned}$$

$$\phi(\tilde{f},\tilde{e}) = \begin{cases} \sum_{\tilde{g}} \phi(\tilde{f},\tilde{g})\ \phi(\tilde{g},\tilde{e}) & \text{integration} \\ \\ \underset{\tilde{g}}{max}\, \phi(\tilde{f},\tilde{g})\ \phi(\tilde{g},\tilde{e}) & \text{maximization} \end{cases}$$

Coupling with Unconstrained Alignments



FONDAZIONE BRUNO KESSLER

Coupling with Unconstrained Alignments

	CE2	CE1	ES1	product disj	product over
src phr	76K	128K	277K	21K	94K
trg phr	82K	134K	284K	32K	108K
phr pairs	133K	185K	333K	592K	696K
avg trans	1.8	1.4	1.2	28.2	7.4
common	-	-	-	59K	143K

	disjoint	overlap
integration	16.65	23.50
maximization	15.88	22.82

- limited intersection among $\tilde{g}$ phrases in the disjoint condition:
  – only 27% of Chinese phrases are bridged into Spanish through English
  – only 11% of Spanish are reached through English
- ambiguity increases (esp. for short phrases)
- integration $>$ maximization
- overlap data would be very useful

Bridging at Training Time

- Standard training criterion for (IBM) alignment models

$$\theta^*_{FE} = \arg\max_{\theta_{FE}} \prod_i p_{\theta_{FE}}(\mathbf{f}_i \mid \mathbf{e}_i) \qquad \text{given } \{(\mathbf{f}_i, \mathbf{e}_i)\}$$

- Goal: estimate parameters of a "direct" F-E system without a (F,E) corpus
- Assumption: a parallel corpus $\{(\mathbf{f}_i, \mathbf{g}_i)\}$, a full-fledged G-E system $p_{\hat{\theta}_{GE}}$
- Solution: $p(\mathbf{f} \mid \mathbf{g})$ above can be replaced with the marginal distribution:

$$\begin{aligned} p(\mathbf{f} \mid \mathbf{g}) &= \sum_{\mathbf{e}} p(\mathbf{f} \mid \mathbf{e})\; p_{\hat{\theta}_{GE}}(\mathbf{e} \mid \mathbf{g}) \\ \hat{\theta}_{FE} &= \arg\max_{\theta_{FE}} \sum_{\mathbf{e}_i} p_{\theta_{FE}}(\mathbf{f}_i \mid \mathbf{e}_i)\; p_{\hat{\theta}_{GE}}(\mathbf{e}_i \mid \mathbf{g}_i) \end{aligned}$$

assuming independence between $\mathbf{e}$ and $\mathbf{f}$ given $\mathbf{g}$.

Approximate ML Estimates

- **Approximation 1**: limit the support of $p_{\hat{\theta}_{GE}}(\mathbf{e} \mid \mathbf{g})$ to the best translation
  – basically, we generate a synthetic parallel corpus (F,$\bar{\text{E}}_F$)
- **Approximation 2**: limit support over the N-best translations
  – requires MLE of IBM models work with two hidden variables
  – still to be developed

We only experimented the first method, called Viterbi approximation

Random Sampling Method

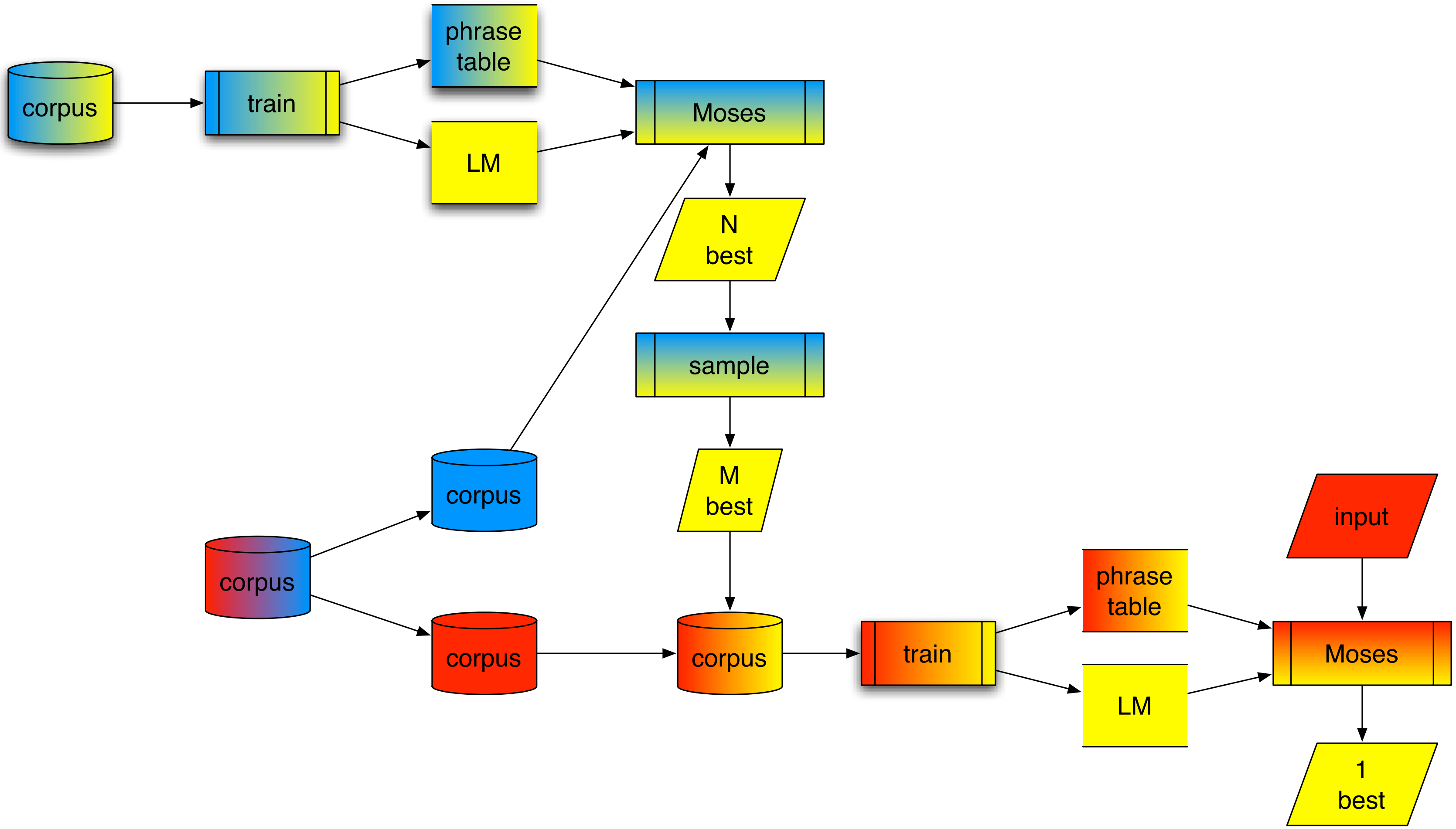
Idea: *Generate parallel data by sampling translations from an SMT system*

- **Ingredients**: corpus (F,G) and SMT system $\mathcal{G} \rightarrow \mathcal{E}$
- For each example $(\mathbf{f}_i, \mathbf{g}_i)$ in the training corpus (F, G) generate a random sample of m translations $\mathbf{e}_{ij}$ of $\mathbf{g}_i$ according to $p_{\hat{\theta}_{GE}}(\mathbf{e} \mid \mathbf{g})$.
- Then build a translation system from $(F, E) = \{(\mathbf{f}_i, \mathbf{e}_{ij})\}, j = 1, \ldots, m$, by maximizing:

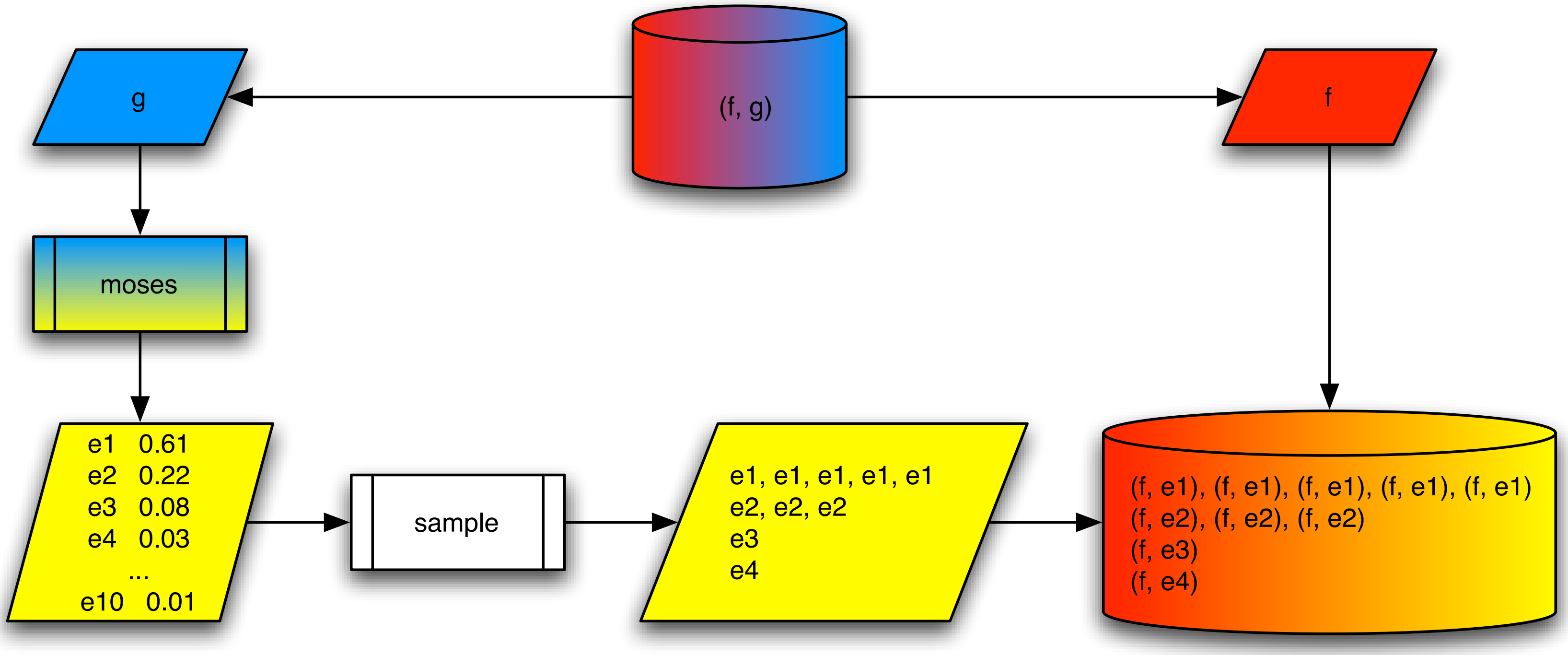
$$\hat{\theta}_{FE} = \arg \underset{\theta_{FE}}{max} \prod_{i,j} P_{\theta_{FE}}(\mathbf{f}_i \mid \mathbf{e}_{ij})$$

- **Implementation**: sample with replacement from the n-best list of translations $\mathbf{e}$ from $\mathbf{g}_i$ according to $p_{\hat{\theta}_{GE}}(\mathbf{e} \mid \mathbf{g}_i)$.
- This approach is indeed more sound than just taking the list of n-best!

Random Sampling Method



Random Sampling Method



- random sampling with replacement 10 times from a 10-best list of translation
- normalization of Moses scores
- more importance to more reliable translations

Random Sampling Method

	n,m	lm	dev	test
Viterbi Training	1	S1	22.05	14.56
Viterbi Training	1	$\bar{\text{S}}$2	23.58	15.38
Viterbi Training	1	S1+$\bar{\text{S}}$2	24.57	16.13
N-best Training	100	S1+$\bar{\text{S}}$2	26.04	17.03
Random Sampling	100	S1+$\bar{\text{S}}$2	26.02	17.68

- LM(E1 $\cup$ $\bar{\text{E}}$2) $>$ LM($\bar{\text{E}}$2) $>$ LM(E1)
- *N-best Training* $>$ *Viterbi Training*
- *N-best Training* $\approx$ *Random Sampling*
- 21% relative improvement wrt Viterbi-S1 (15% wrt Viterbi-$\bar{\text{S}}$2)

Experimental Results

	CES task	
Method	disjoint	overlap
Direct	–	23.67
Cascade 1-best	16.44	24.04
Cascade N-best	17.64	25.16
PhraseTable Product	16.65	23.50
Random Sampling	17.68	25.19

- *Cascade 1-best* $\approx$ *PhraseTable Product*
- *Random Sampling* $\approx$ *Cascade N-best* $>$ *Direct*

Summary

- approaches to pivot translation task
- mathematical foundation
- experimental comparison

- random sampling approach is the most appealing:
  – quality and efficiency

- unsupervised technique to generate new parallel data
  – suitable to domain adaptation
  – suitable for multi-language pivot translation

Thank you!