FONDAZIONE
BRUNO KESSLER

FBK @ IWSLT-2008

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

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

October 20th, 2008

FONDAZIONE BRUNO KESSLER

FBK goal

Pivot translation in real-world condition

- improving translation for low-resourced languages:
  - few parallel data for Italian-centric language pairs: Chinese, Arabic, ...
- improving translation among intra-European languages
- applying pivot-like strategies to adapt SMT systems to different domains

- theoretical foundation of pivot translation task
- mathematically sound definition of approaches
- experimental comparison

FBK @ IWSLT 2008

Most effort on Pivot Task

- good benchmark:
  - controlled conditions, controlled domain
  - fast development cycle because of small size
  - many competitors

- participation to other IWSLT tasks, but with limited effort:
  - no use of additional data
  - no adaptation to challenge task
  - no optimization for speech input

FONDAZIONE BRUNO KESSLER

Task Description

- traveling domain
- Basic Travel Expression Corpus

- BTEC tasks:
  – translation from Chinese into English and from Chinese into Spanish
- **Pivot task**:
  – translation from Chinese into Spanish without C-S parallel data
  – only *independent* C-E and E-S parallel data available
- Challenge task:
  – translation from Chinese into English of tourism-related dialogues (no BTEC)

- input condition:
  – automatic and **correct transcriptions**
  – read (BTEC and Pivot) and spontaneous (Challenge) speech

FONDAZIONE BRUNO KESSLER

Task description: data

- training data:
  - monolingual corpora: C1 and C2, E1 and E2, and S1
  - parallel corpora: CE2, ES1, development sets (with multiple refs)
  - CES1 never used as trilingual parallel corpus
  - **no additional data** (although allowed)

- development data
  - dev set: 506 Chinese sentences with 16 refs in English and Spanish
  - other dev sets for C-E BTEC and Challenge tasks
  - **blind devtest** set: 1K sentences with 1 reference
  - **reduced training** corpora (19K sentences) for development

- test set: 507 Chinese sentences

- preprocessing: tokenization, numbers into digits, Chinese word-segmentation

Pivot Task description: data

task	data	sent	source		target	
			words	dict	words	dict
Btec	CE1*	18,974	161K	8,017	172K	8,210
	CS1*	18,974	161K	8,017	176K	10,773
Pivot	CE2*	18,999	150K	8,114	172K	8,631
	ES1*	18,974	172K	8,210	176K	10,773
Btec	CE1+dev	54,021	439K	8,847	499K	10,765
	CS1+dev	28,068	229K	8,284	250K	11,734
Pivot	CE2+dev	28,095	217K	8,987	248K	8,951
	ES1+dev	19,972	182K	8,385	177K	11,019
Challenge	CE1+dev	55,743	447K	8,864	507K	11,051

- training data during development (*)
- training data the final submissions including development sets (+dev)

FONDAZIONE BRUNO KESSLER

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

Direct system: performance

	data	BLEU	OOV	applied to
Chinese-English	CE1*	26.91	2.00	Btec and Challenge
	CE2*	19.09	3.80	Pivot
English-Spanish	ES1*	49.13	2.01	Pivot
Chinese-Spanish	CS1*	23.67	2.00	Btec

- systems trained on reduced data
- performance on the blind devtest, extracted from CE1 and ES1

- significant mismatch between corpora 1 and 2
- translation from Chinese into English easier than into Spanish
- translation from English into Spanish "easy"

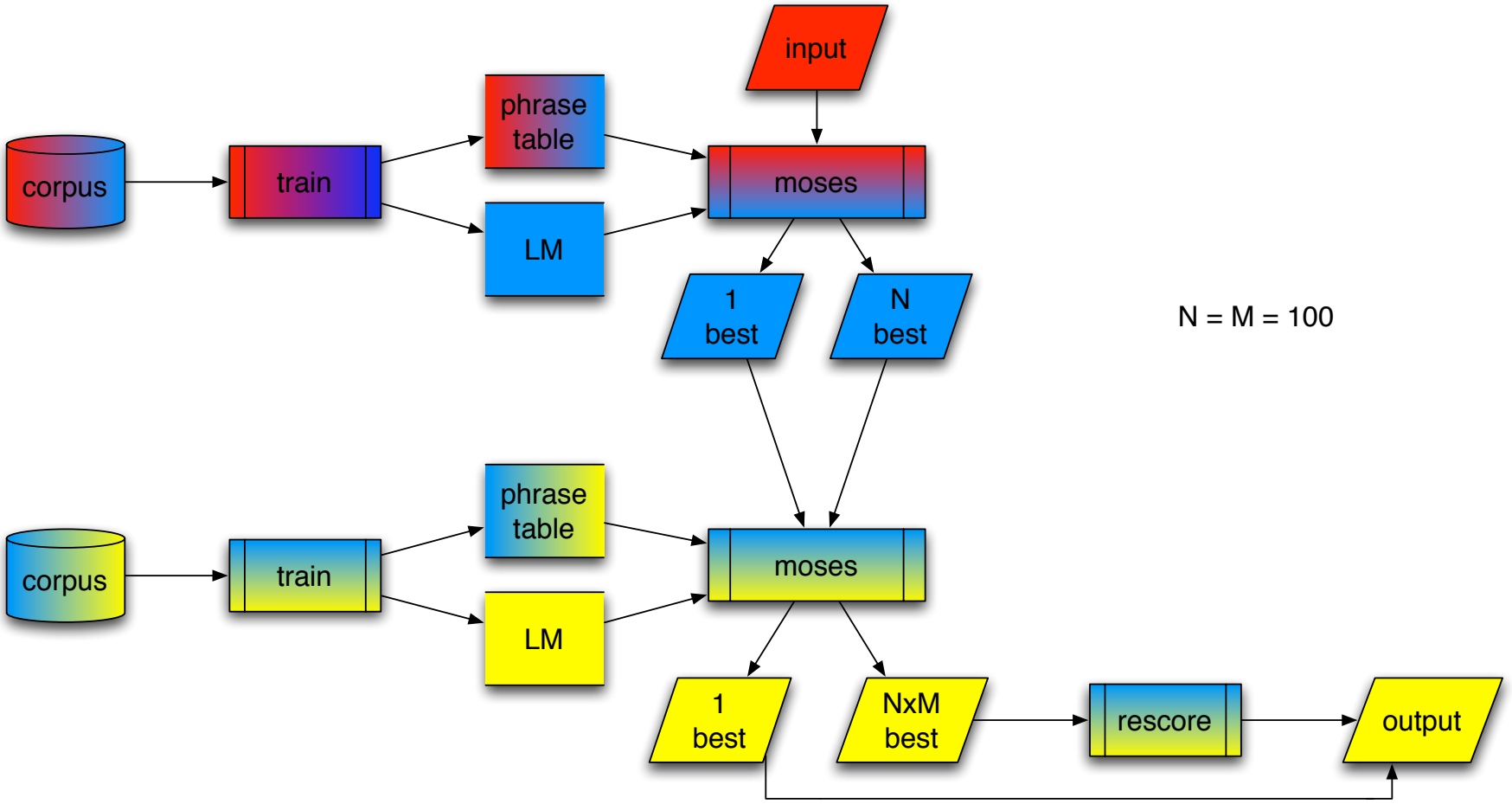
Pivot SMT

- Goal:
  - translation from Chinese into Spanish without parallel data

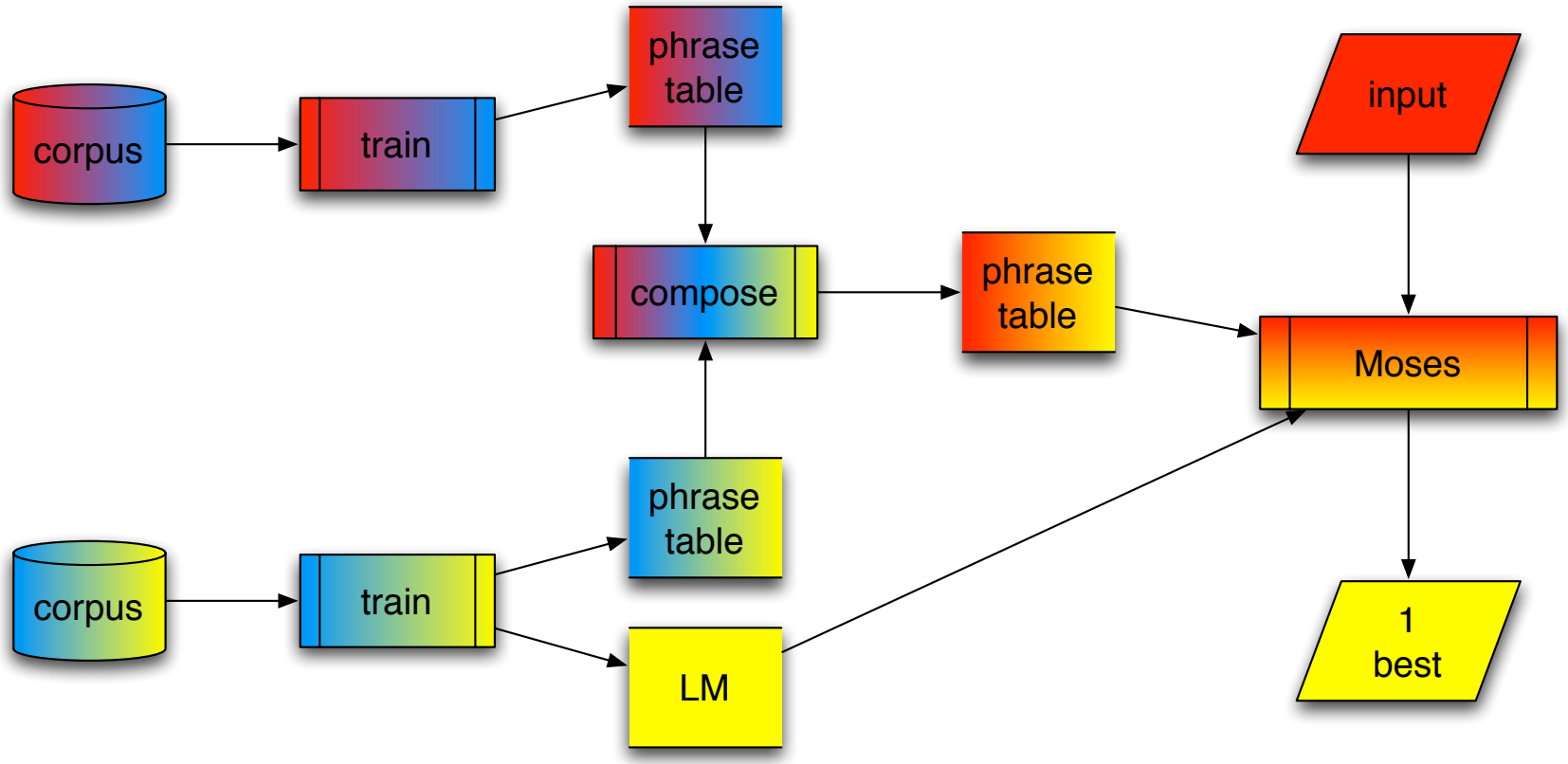
- Assumption:
  - two parallel corpora C-E and E-S, with **independent** English side
  - **full-fledged** *Direct* systems trained on C-E and E-S parallel data

- Approaches:
  - Coupling C-E and E-S systems at **sentence level**
  - Coupling C-E and E-S systems at **phrase level**
  - **Synthesizing** C-S parallel data and building a full-fledged C-S system

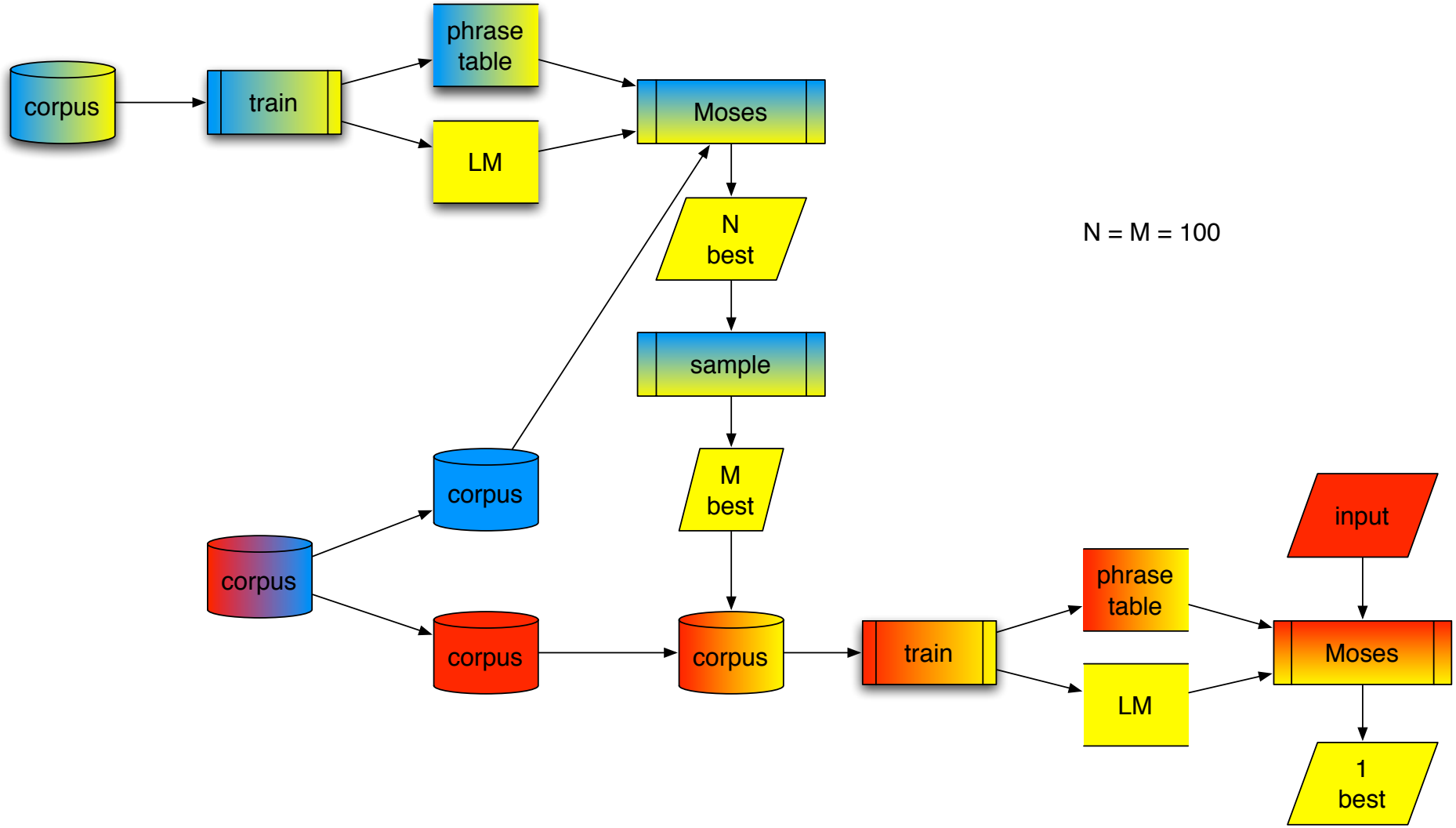
Coupling systems at sentence level



Coupling systems at phrase level



Synthesis of parallel data



FONDAZIONE BRUNO KESSLER

Official results of Pivot Task

system	run	ASR.1	CRR
Cascade 1-best	contr6	29.20	33.52
Cascade Nbest	contr7	32.69	37.41
PT Composition	contr4	28.52	33.13
Synthesis	**prim**	**33.11**	**39.69**
	contr1	34.14	39.93

- big gain using 100-best wrt to 1best
- less than 2 BLEU points wrt top performing (39.69 vs 41.57)
- avoiding the CE translation, which poorly performs, is a winning strategy
- ASR (- 13/17% relative) confirms the same results as CRR

- contr1 includes the C-S parallel data of the dev set, **not independent data**
- using correct Spanish translations is better than using synthesized ones

Thank you!

FONDAZIONE BRUNO KESSLER

Official results of all submissions

Task	System	Run	BLEU	
			ASR.1	CRR
CE-btec	*Direct*	**prim**	**36.91**	**40.18**
		contr	36.45	"
CS-btec	*Direct*	**prim**	**26.67**	**30.29**
		contr	27.05	"
CE-chal	*Direct*	**prim**	**23.84**	**27.00**
		contr	23.88	"
CES-pivot	*Cascade*	contr6	29.20	33.52
	Nbest	contr7	32.69	37.41
	PhraseTable	contr4	28.52	33.13
		contr5	30.09	"
	Synthesis	**prim**	**33.11**	**39.69**
		contr2	35.94	'
		contr1	34.14	39.93
		contr3	35.98	"