Improving the quality of a customized SMT system using shared training data

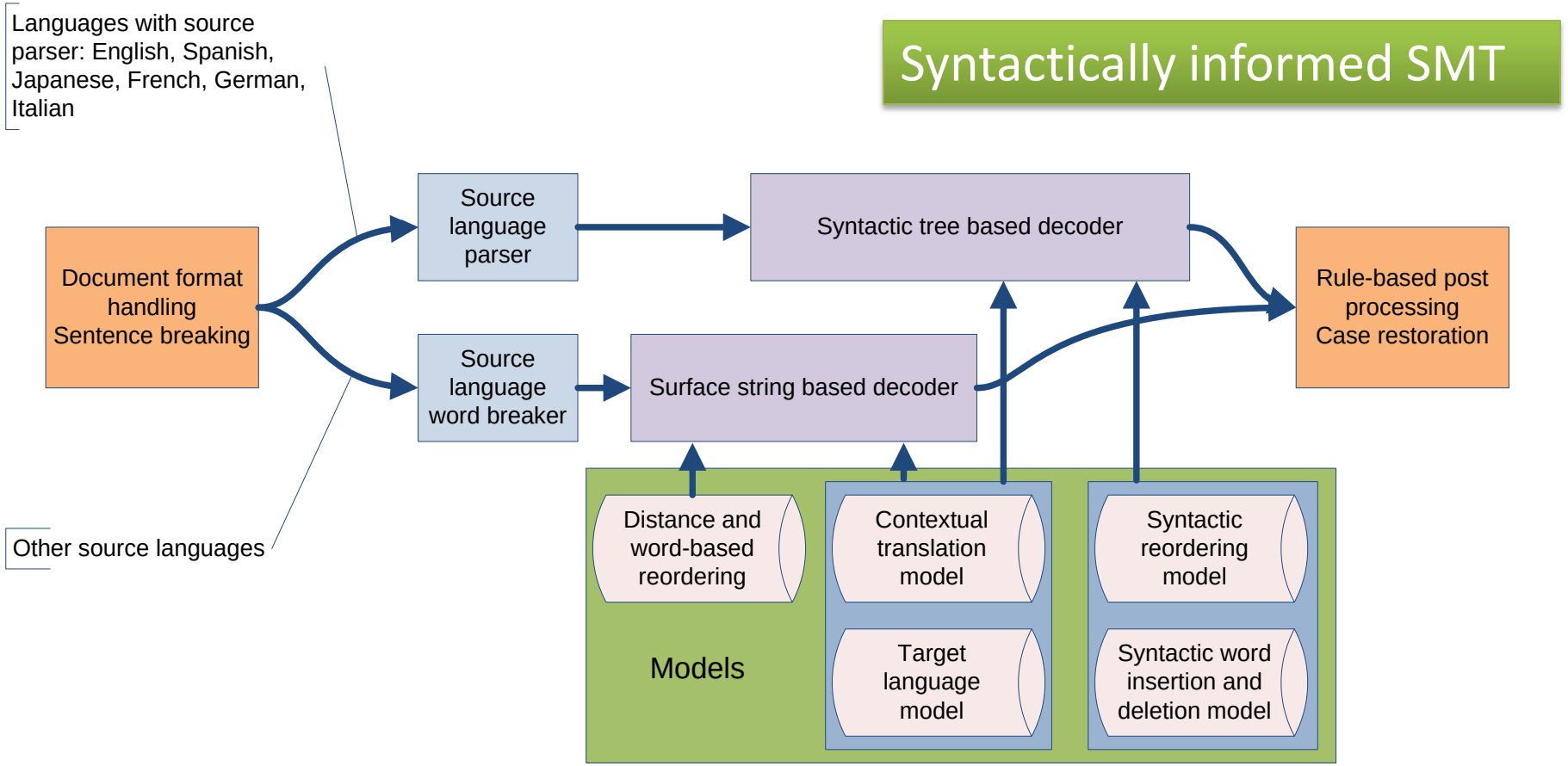
Chris.Wendt@microsoft.com
Will.Lewis@microsoft.com
August 28, 2009

Microsoft® Translator

Overview

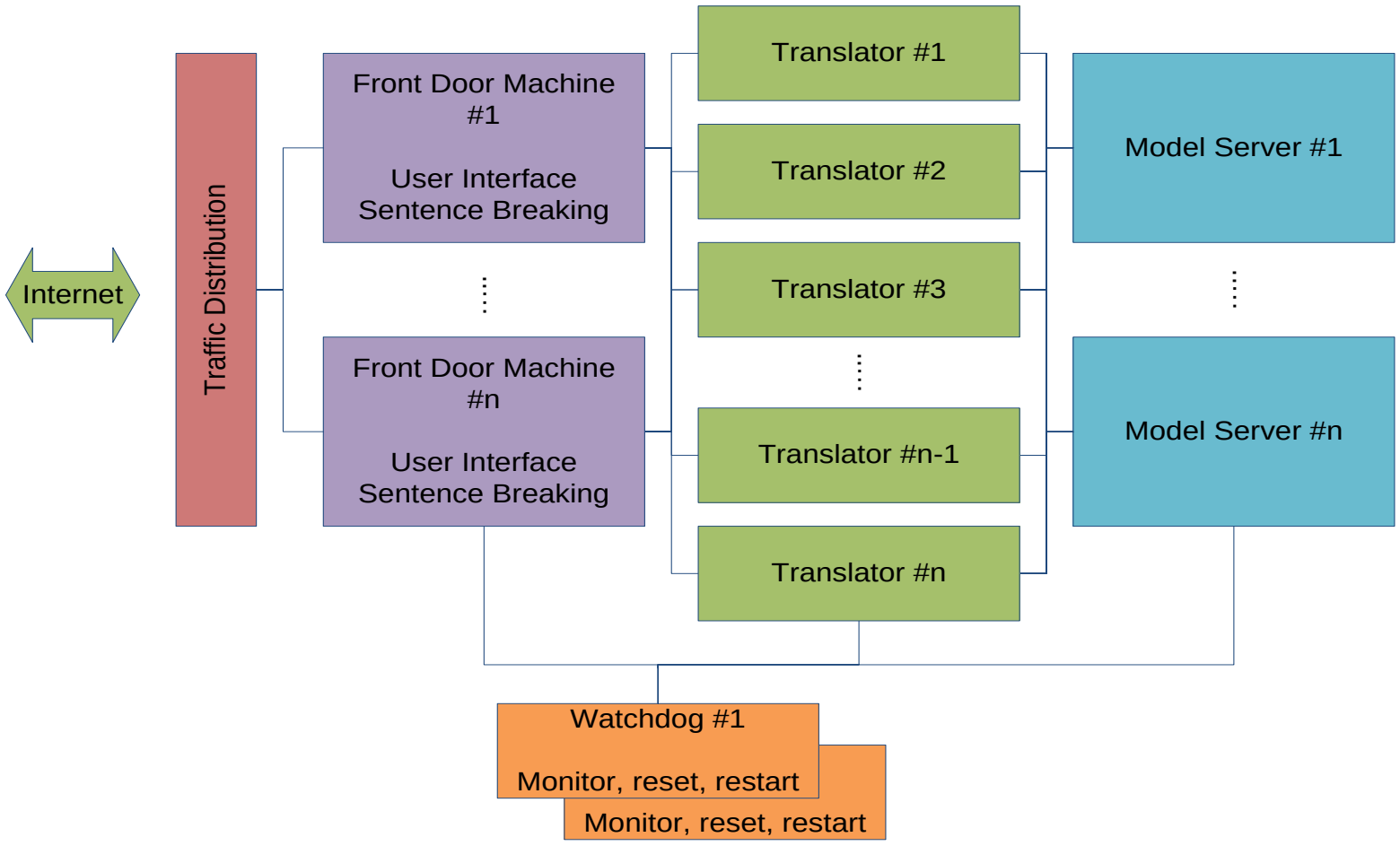
- Engine and Customization Basics
- Experiment Objective
- Experiment Setup
- Experiment Results
- Validation
- Conclusions

Microsoft's Statistical MT Engine



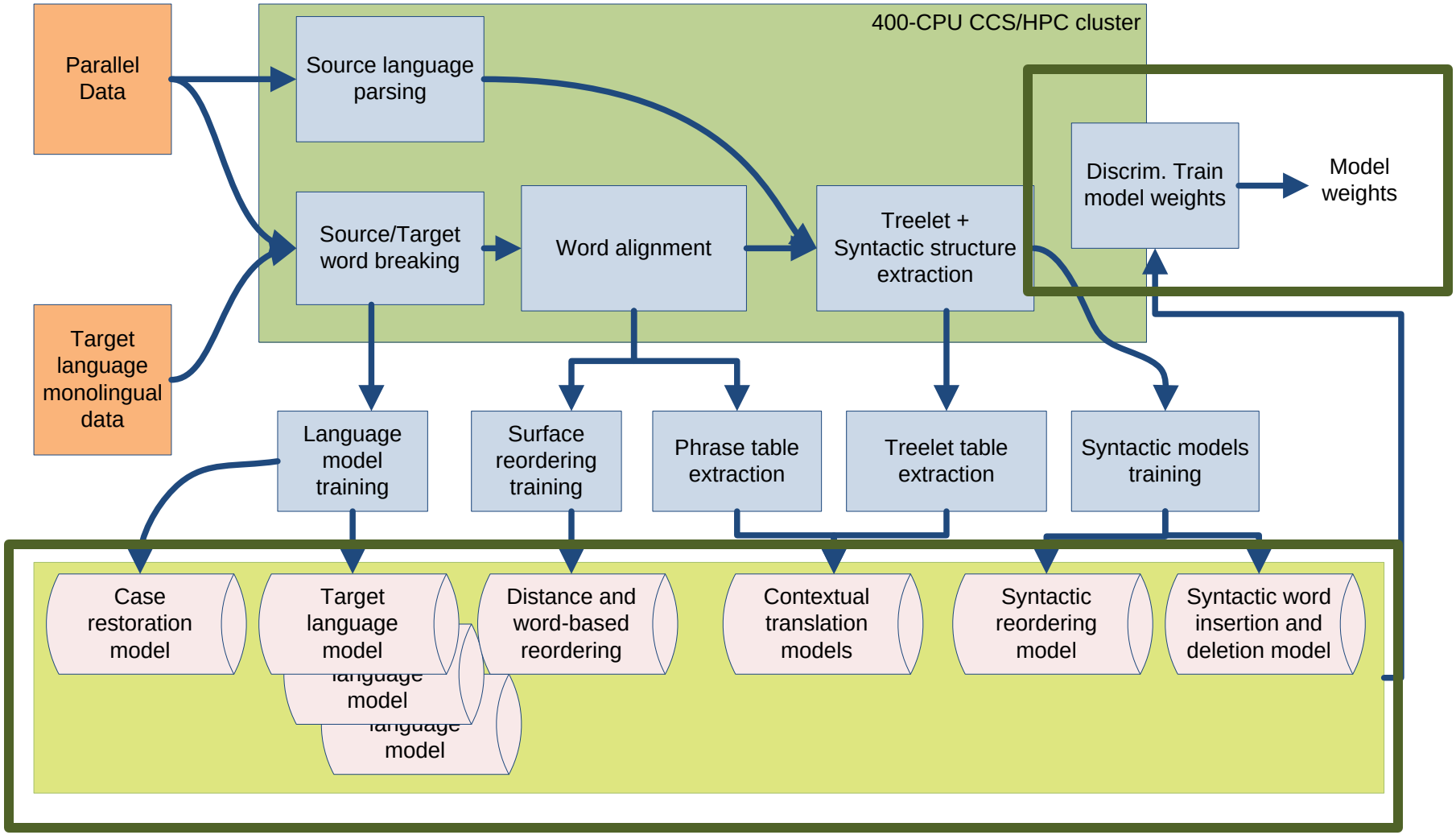
Microsoft® Translator

Microsoft Translator Runtime



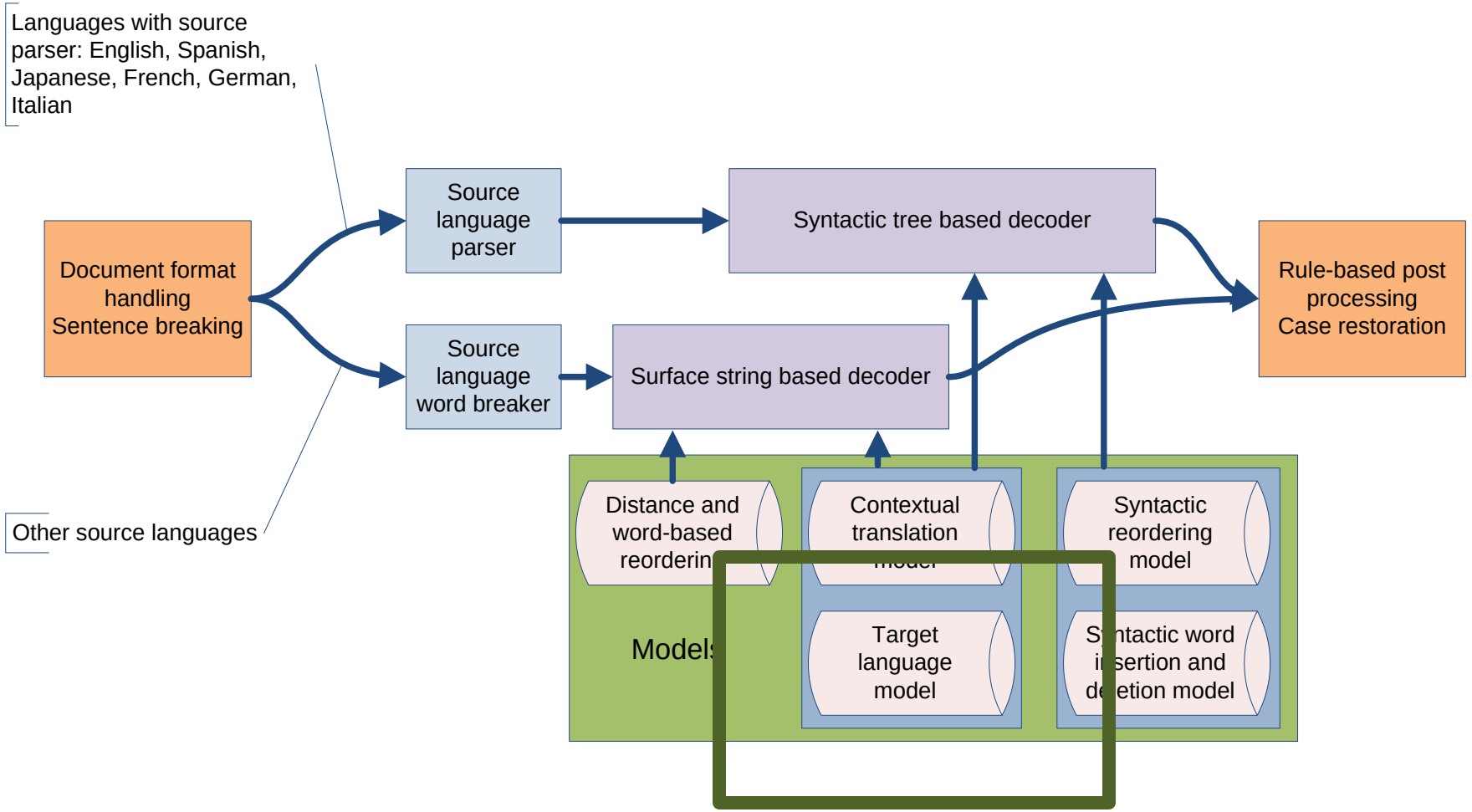
Microsoft® Translator

Training



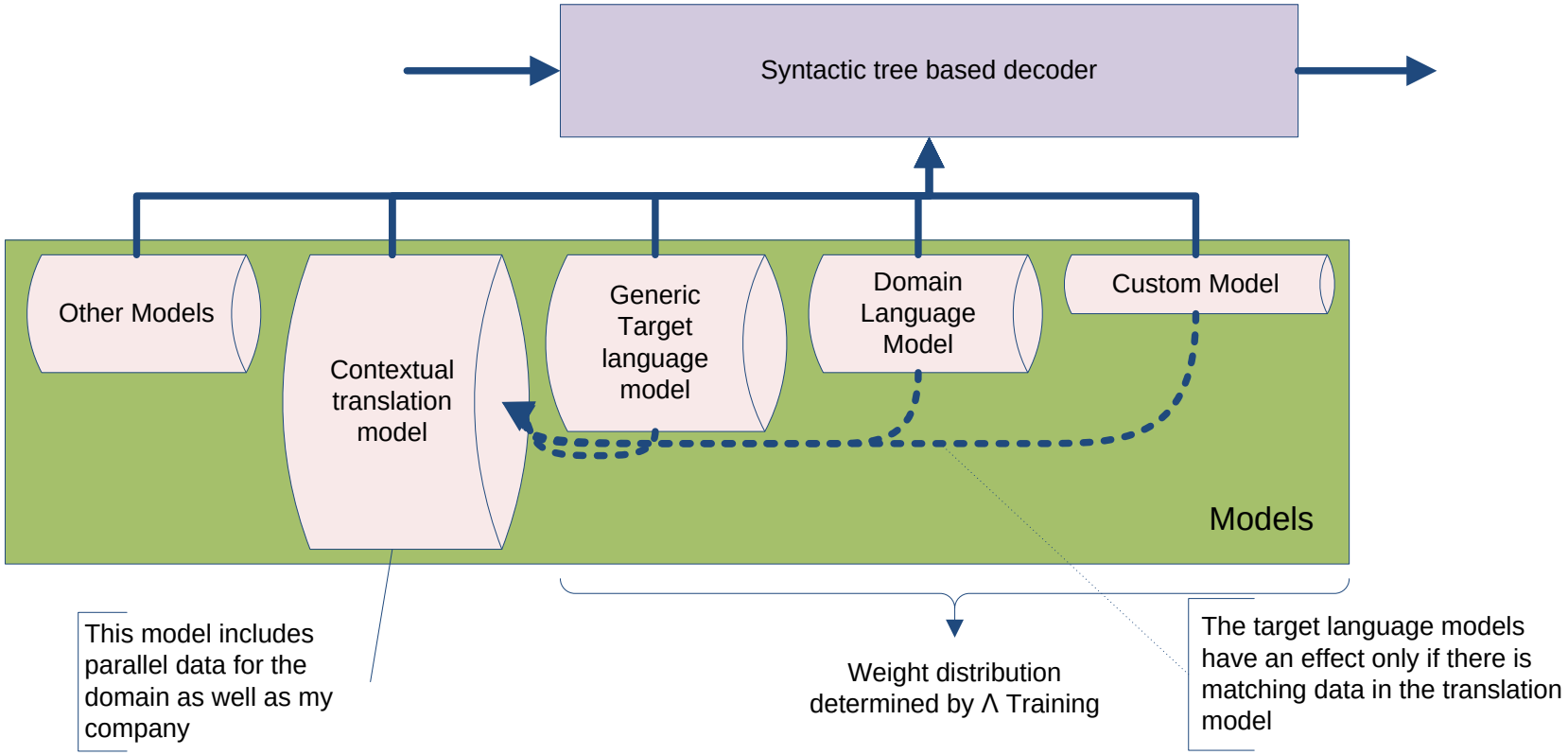
Microsoft® Translator

Microsoft's Statistical MT Engine



Microsoft® Translator

Adding Domain Specificity



Experiment Objective

Objective

- Determine the effect of pooling parallel data among multiple data providers within a domain, measured by the translation quality of an SMT system trained with that data.

Experiment Setup

1. Data pool: TAUS Data Association's repository of parallel translation data.
2. Domain: computer-related technical documents.

   No difference is made between software, hardware, documentation and marketing material.
3. Criteria for test case selection:
   - More than 100,000 segments of parallel training data
   - Less than 2M segments of parallel training data (at that point it would be valid to train a System using *only* the provider's own data)
4. Chosen case: Sybase
5. Experiment Series: Observe BLEU scores using a reserved subset of the submitted data against systems trained with
   - 1 General data, as used for www.microsofttranslator.com
   - 2a Only Microsoft's internal parallel data, from localization of its own products
   - 2b Microsoft data + Sybase data
   - 3a General + Microsoft + TAUS
   - 3b General + Microsoft data + TAUS, with Sybase custom lambdas
6. Measure BLEU on 3 sets of test documents, with 1 reference, reserved from the submission, not used in training:
   - Sybase
   - Microsoft
   - General

System Details

ID	Parallel Data	Target Language Models	Lambda
1	General	General	General
2a	Microsoft	Microsoft	Microsoft
2b	Microsoft and Sybase	Microsoft and Sybase	Sybase
3a	General and Microsoft and TAUS	General Microsoft and TAUS	TAUS
3b	General and Microsoft and TAUS	General Microsoft and TAUS Sybase	Sybase

Training data composition

Chinese (Simplified)

Classification	Provider	Segments
Hardware	Intel	281903
Hardware	EMC	757142
Hardware	Dell	347945
Software	EMC	103862
Software	McAfee	213790
Software	Sybase iAnywhere	240389
Software	Avocent	81348
Software	Sun Microsystems	183498
Software	Adobe	153670
Software	PTC	142965
Software	Intel	259
Software	SDL	25064
Software	Microsoft	5029554

German

Classification	Provider	Segments
Hardware	EMC	414791
Hardware	Intel	128209
Hardware	Dell	314496
Professional	eBay, Inc.	59967
Software	Avocent	93498
Software	EMC	124065
Software	McAfee	497938
Software	Sybase iAnywhere	216315
Software	ABBYY	28063
Software	Adobe	232914
Software	Sun Microsystems	51644
Software	PTC	178341
Software	Intel	11566
Software	SDL	44029
Software	Microsoft	6172394

Sybase does not have enough data to build a system exclusively with Sybase data

Microsoft® Translator

Experiment Results, measured in BLEU

Chinese

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	8.3M	General domain	14.26	29.74	34.81
2a	2.6M	Microsoft	12.32	34.65	29.95
2b	2.8M	Microsoft with Sybase	12.16	34.66	30.24
3a	11.5M	General and Microsoft and TAUS	15.38	35.80	44.49
3b	11.5M	System 3a with Sybase lambda	12.57	29.51	47.16

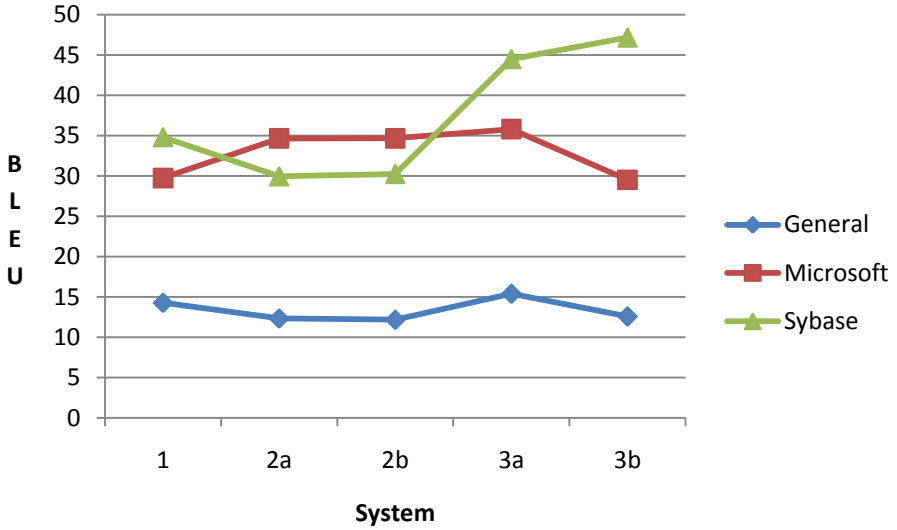
German

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	4.4M	General Domain	25.19	40.61	34.85
2a	7.6M	Microsoft	21.95	52.39	41.55
2b	7.8M	Microsoft with Sybase	22.83	52.07	42.07
3a	11.1M	General and Microsoft and TAUS	23.86	52.72	48.83
3b	11.1M	System 3a with Sybase lambda	19.44	37.27	50.85

Experiment Results, measured in BLEU

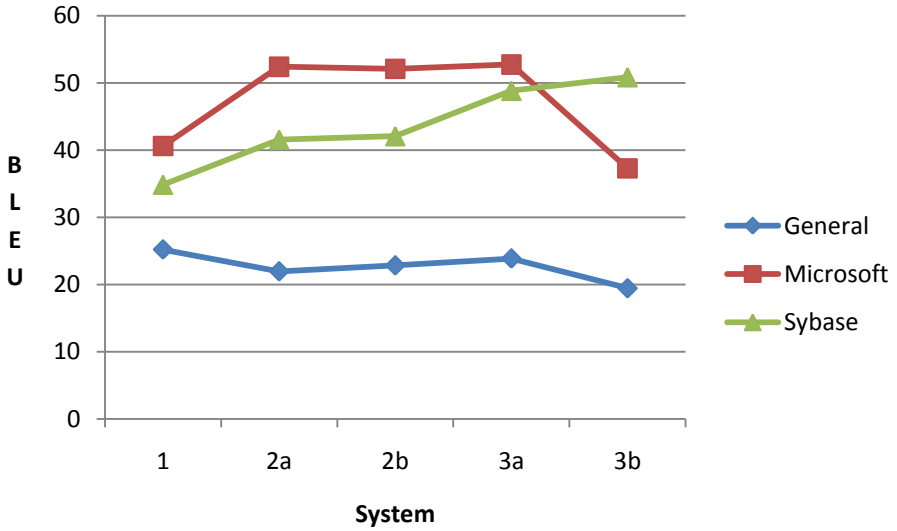
Chinese

System	Size	System Description
1	8.3M	General domain
2a	2.6M	Microsoft
2b	2.8M	Microsoft with Sybase
3a	11.5M	General and Microsoft and T
3b	11.5M	System 3a with Sybase lamb



German

System	Size	System Description
1	4.4M	General Domain
2a	7.6M	Microsoft
2b	7.8M	Microsoft with Sybase
3a	11.1M	General and Microsoft and T
3b	11.1M	System 3a with Sybase lamb



Microsoft® Translator

Experiment Results, measured in BLEU

Chinese

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	8.3M	General domain	14.26	29.74	34.81
2a	2.6M	Microsoft	12.32	34.65	29.95
2b	2.8M	Microsoft with Sybase	12.16	34.66	30.24
3a	11.5M	General and Microsoft and TAUS	15.38	35.80	44.49
3b	11.5M	System 3a with Sybase lambda	12.57	29.51	47.16

German

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	4.4M	General Domain	25.19	40.61	34.85
2a	7.6M	Microsoft	21.95	52.39	41.55
2b	7.8M	Microsoft with Sybase	22.83	52.07	42.07
3a	11.1M	General and Microsoft and TAUS	23.86	52.72	48.83
3b	11.1M	System 3a with Sybase lambda	19.44	37.27	50.85

Experiment Results, measured in BLEU

Chinese

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	8.3M	General domain	14.26	29.74	34.81
2a	2.6M	Microsoft	12.32	34.65	29.95
2b	2.8M	Microsoft with Sybase	12.16	34.66	30.24
3a	11.5M	General and Microsoft and TAUS	15.38	35.80	44.49
3b	11.5M	System 3a with Sybase lambda	12.57	29.51	47.16

Germa

More than 8 point gain compared to system built without the shared data

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	4.4M	General Domain	25.19	40.61	34.85
2a	7.6M	Microsoft	21.95	52.39	41.55
2b	7.8M	Microsoft with Sybase	22.83	52.07	42.07
3a	11.1M	General and Microsoft and TAUS	23.86	52.72	48.83
3b	11.1M	System 3a with Sybase lambda	19.44	37.27	50.85

Experiment Results, measured in BLEU

Chinese

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	8.3M	General domain	14.26	29.74	34.81
2a	2.6M	Microsoft	12.32	34.65	29.95
2b	2.8M	Microsoft with Sybase	12.16	34.66	30.24
3a	11.5M	General and Microsoft and TAUS	15.38	35.80	44.49
3b	11.5M	System 3a with Sybase lambda	12.57	29.51	47.16

Best results are achieved using the maximum available data within the domain, using custom lambda training

			Test Set		
			General	**Microsoft**	**Sybase**
1	4.4M	General Domain	25.19	40.61	34.85
2a	7.6M	Microsoft	21.95	52.39	41.55
2b	7.8M	Microsoft with Sybase	22.83	52.07	42.07
3a	11.1M	General and Microsoft and TAUS	23.86	52.72	48.83
3b	11.1M	System 3a with Sybase lambda	19.44	37.27	50.85

Experiment Results, measured in BLEU

Chinese

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	8.3M	General domain	14.26	29.74	34.81
2a	2.6M	Microsoft	12.32	34.65	29.95
2b	2.8M	Microsoft with Sybase	12.16	34.66	30.24
3a	11.5M	General and Microsoft and TAUS	15.38	35.80	44.49
3b	11.5M	System 3a with Sybase lambda	12.57	29.51	47.16

Weight training (lambda training) without diversity in the training data has very little effect

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	4.4M	General Domain	25.19	40.61	34.85
2a	7.6M	Microsoft	21.95	52.39	41.55
2b	7.8M	Microsoft with Sybase	22.83	52.07	42.07
3a	11.1M	General and Microsoft and TAUS	23.86	52.72	48.83
3b	11.1M	System 3a with Sybase lambda	19.44	37.27	50.85

Microsoft® Translator

The diversity aspect was somewhat a surprise for us. Microsoft's large data pool by itself did not give Sybase the hoped-for boost.

Experiment Results, measured in BLEU

Chinese

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	8.3M	General domain	14.26	29.74	34.81
2a	2.6M	Microsoft	12.32	34.65	29.95
2b	2.8M	Microsoft with Sybase	12.16	34.66	30.24
3a	11.5M	General and Microsoft and TAUS	15.38	35.80	44.49
3b	11.5M	System 3a with Sybase lambda	12.57	29.51	47.16

Lambda training with in-domain diversity has a significant positive effect for the lambda target, and a significant negative effect for everyone else

			Test Set		
			General	**Microsoft**	**Sybase**
1	4.4M	General Domain	25.19	40.61	34.85
2a	7.6M	Microsoft	21.95	52.39	41.55
2b	7.8M	Microsoft with Sybase	22.83	52.07	42.07
3a	11.1M	General and Microsoft and TAUS	23.86	52.72	48.83
3b	11.1M	System 3a with Sybase lambda	19.44	37.27	50.85

Experiment Results, measured in BLEU

Chinese

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	8.3M	General domain	14.26	29.74	34.81
2a	2.6M	Microsoft	12.32	34.65	29.95
2b	2.8M	Microsoft with Sybase	12.16	34.66	30.24
3a	11.5M	General and Microsoft and TAUS	15.38	35.80	44.49
3b	11.5M	System 3a with Sybase lambda	12.57	29.51	47.16

A system can be customized with small amounts of target language material, as long as there is a diverse set of in-domain parallel data available

			Test Set		
			General	**Microsoft**	**Sybase**
1	4.4M	General Domain	25.19	40.61	34.85
2a	7.6M	Microsoft	21.95	52.39	41.55
2b	7.8M	Microsoft with Sybase	22.83	52.07	42.07
3a	11.1M	General and Microsoft and TAUS	23.86	52.72	48.83
3b	11.1M	System 3a with Sybase lambda	19.44	37.27	50.85

Experiment Results, measured in BLEU

Chinese

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	8.3M	General domain	14.26	29.74	34.81
2a	2.6M	Microsoft	12.32	34.65	29.95
2b	2.8M	Microsoft with Sybase	12.16	34.66	30.24
3a	11.5M	General and Microsoft and TAUS	15.38	35.80	44.49
3b	11.5M	System 3a with Sybase lambda	12.57	29.51	47.16

Small data providers benefit more from sharing than large data providers, but all benefit

			Test Set		
			General	**Microsoft**	**Sybase**
1	4.4M	General Domain	25.19	40.61	34.85
2a	7.6M	Microsoft	21.95	52.39	41.55
2b	7.8M	Microsoft with Sybase	22.83	52.07	42.07
3a	11.1M	General and Microsoft and TAUS	23.86	52.72	48.83
3b	11.1M	System 3a with Sybase lambda	19.44	37.27	50.85

Experiment Results, measured in BLEU

Chinese

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	8.3M	General domain	14.26	29.74	34.81
2a	2.6M	Microsoft	12.32	34.65	29.95
2b	2.8M	Microsoft with Sybase	12.16	34.66	30.24
3a	11.5M	General and Microsoft and TAUS	15.38	35.80	44.49
3b	11.5M	System 3a with Sybase lambda	12.57	29.51	47.16

This is the best German Sybase system we could have built without TAUS

			Test Set		
System	**Size**	**System Description**	**General**	**Microsoft**	**Sybase**
1	4.4M	General Domain	25.19	40.61	34.85
2a	7.6M	Microsoft	21.95	52.39	41.55
2b	7.8M	Microsoft with Sybase	22.83	52.07	42.07
3a	11.1M	General and Microsoft and TAUS	23.86	52.72	48.83
3b	11.1M	System 3a with Sybase lambda	19.44	37.27	50.85

Microsoft® Translator

Validation: Adobe Polish

		Test Set		
System Size	**System Description**	**General**	**Microsoft**	**Adobe**
1	General domain	15.90	28.90	19.40
2a	Microsoft			
2b	Microsoft with Adobe			
3a	General and Microsoft and TAUS			
3b	System 3a with Adobe lambda	13.53	33.88	33.74

Training Data (sentences):

- General 1.5M
- Microsoft 1.7M
- Adobe 129K
- TAUS other 70K

Even for a language without a lot of training data we can see nice gains by pooling.

Validation: Dell Japanese

		Test Set		
System Size	**System Description**	**General**	**Microsoft**	**Dell**
1	General domain	17.99	37.88	26.72
2a	Microsoft	17.28	41.32	32.64
2b	Microsoft with Dell	14.76	30.87	39.49
3a	General and Microsoft and TAUS	17.33	42.30	39.89
3b	System 3a with Dell lambda	14.85	32.21	42.43

Training data (sentences)

- General 4.3M
- Microsoft 3.2M
- TAUS 1.4M
- Dell 172K

Confirms the Sybase results

Example

SRC	The Monitor collects metrics and performance data from the databases and MobiLink servers running on other computers, while a separate computer accesses the Monitor via a web browser.
1	Der Monitor sammelt Metriken und Leistungsdaten von Datenbanken und MobiLink-Servern, die auf anderen Computern ausführen, während auf ein separater Computer greift auf den Monitor über einen Web-Browser.
2a	Der Monitor sammelt Metriken und **Performance-Daten** von **der** Datenbanken und MobiLink-Serve**r** auf anderen Computern **ausgeführt werden**, während **ein** separater Computer den Monitor über einen **Webbrowser** zugreift.
2b	Der Monitor sammelt Metriken und Performance-Daten von der Datenbanken und MobiLink-Server auf anderen Computern ausgeführt werden, während ein separater Computer den Monitor über einen Webbrowser zugreift.
3a	Der Monitor sammelt Metriken und Performance-Daten von der Datenbanken und MobiLink-Server auf anderen Computern ausgeführt werden, während ein separater Computer den Monitor über einen Webbrowser zugreift.
3b	Der Monitor sammelt **Kriterien** und Performance-Daten **aus** der Datenbanken und MobiLink-Server auf anderen Computern ausgeführt werden, während ein separater Computer **des** Monitors über einen Webbrowser zugreift.
REF	Der Monitor sammelt Kriterien und Performance-Daten aus de**n** Datenbanken und MobiLink-Server**n die** auf anderen Computern ausgeführt werden, während ein separater Computer **auf** de**n** Monito**r** über einen Webbrowser zugreift.
Google	Der Monitor sammelt Metriken und Performance-Daten aus den Datenbanken und MobiLink-Server auf anderen Computern ausgeführt, während eine separate Computer auf dem Monitor über einen Web-Browser.

Microsoft® Translator

Bold signals delta to predecessor.

Observations

- Combining in-domain training data gives a significant boost to MT quality. In our experiment more than 8 BLEU points compared to the best System built without the shared data.
- Weight training (Lambda training) without diversity in the training data has almost no effect
- Lambda training with in-domain diversity has a significant positive effect for the lambda target, and a significant negative effect for everyone else
- A system can be customized with small amounts of target language material, as long as there is a diverse set of in-domain parallel data available
- Best results are achieved using the maximum available data within the domain, using custom lambda training
- Small data providers benefit more from sharing than large data providers, but all benefit

Results

- There is noticeable benefit in sharing parallel data among multiple data owners within the same domain, as is the intent of the TAUS Data Association.
- An MT system trained with the combined data can deliver significantly improved translation quality, compared to a system trained only with the provider's own data plus baseline training.
- Customization via a separate target language model and lambda training works

References

- Chris Quirk, Arul Menezes, and Colin Cherry, Dependency Treelet Translation: Syntactically Informed Phrasal SMT, in *Proceedings of ACL, Association for Computational Linguistics*, June 2005
- Microsoft Translator: www.microsofttranslator.com
- TAUS Data Association: www.tausdata.org