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

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

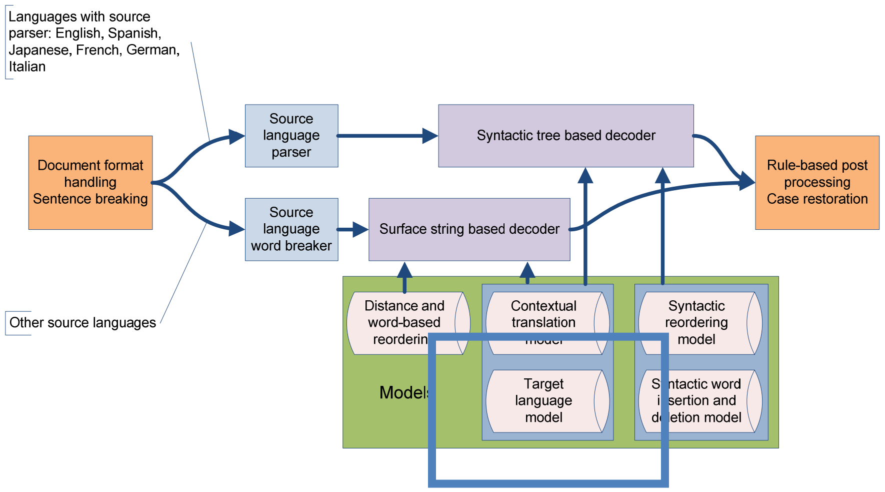
Microsoft Translator

Microsoft Translator - Overview

- Engine and Customization Basics
- Objective
- Experiment setup
- Experiment results

Microsoft Translator

Microsoft's Statistical MT Engine



Languages with source parser: English, Spanish, Japanese, French, German, Italian

Other source languages

Document format handling Sentence breaking

Source language parser

Source language word breaker

Syntactic tree based decoder

Surface string based decoder

Rule-based post processing Case restoration

Distance and word-based reorderin

Contextual translation model

Syntactic reordering model

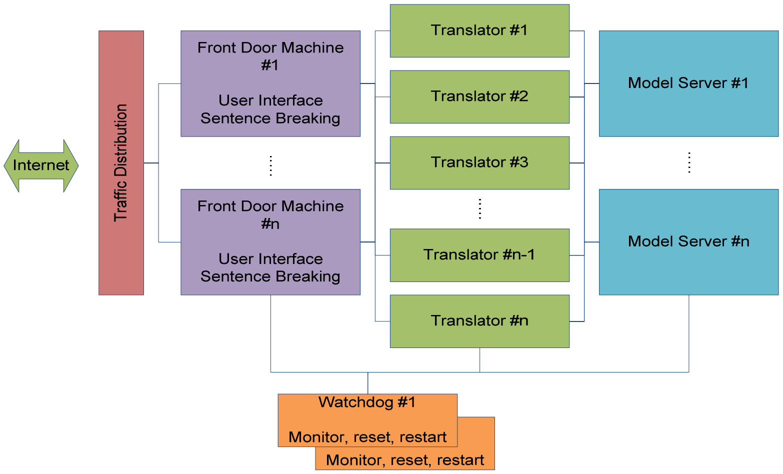
Models

Target language model

S ntactic word i sertion and d letion model

Microsoft Translator

Microsoft Translator Runtime



Internet

Traffic Distribution

Front Door Machine #1

User Interface Sentence Breaking

Front Door Machine #n

User Interface Sentence Breaking

Translator #1

Translator #2

Translator #3

Translator #n-1

Translator #n

Model Server #1

Model Server #n

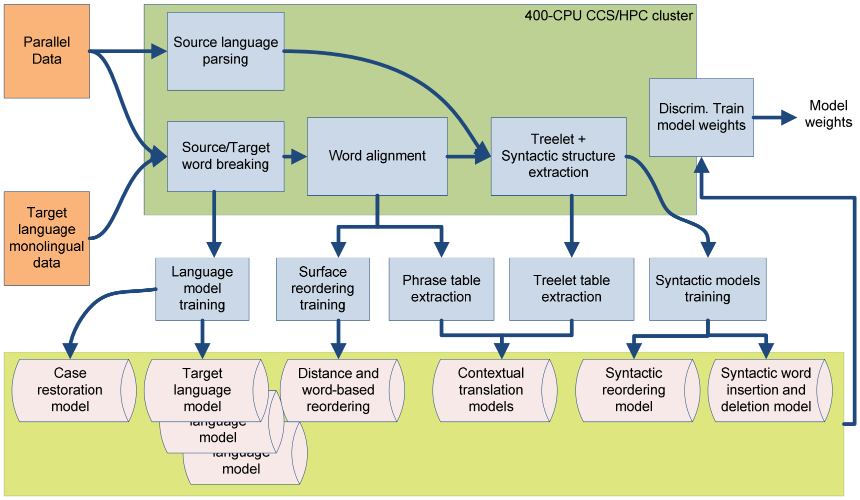
Watchdog #1

Monitor, reset, restart

Monitor, reset, restart

Microsoft Translator

Training



400-CPU CCS/HPC cluster

Parallel Data

Target language monolingual data

Source language parsing

Source/Target word breaking

Word alignment

Treelet + Syntactic structure extraction

Discrim. Train model weights

Model weights

Language model training

Surface reordering training

Phrase table extraction

Treelet table extraction

Syntactic models training

Case restoration model

Target language model

language model

language model

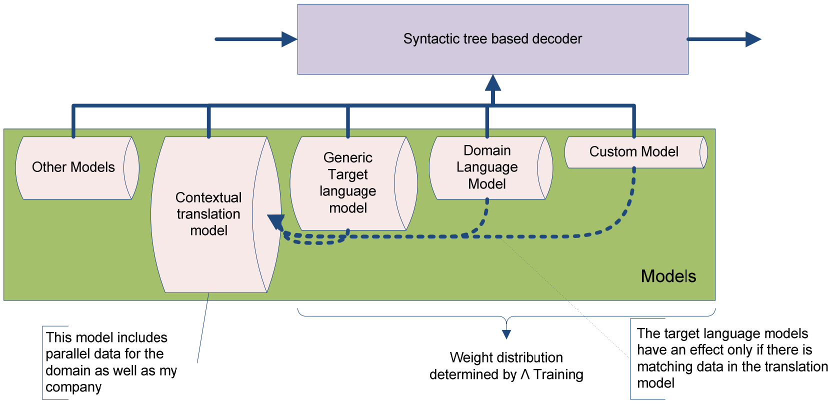
Distance and word-based reordering

Contextual translation models

Syntactic reordering model

Syntactic word insertion and deletion model

Adding Domain Specificity



Syntactic tree based decoder

Other Models

Contextual translation model

Generic Target language model

Domain Language Model

Custom Model

Models

This model includes parallel data for the domain as well as my company

Weight distribution determined by Λ Training

The target language models have an effect only if there is matching data in the translation model

Objective and Result

Objective

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

Result

- There is noticeable benefit in sharing parallel data among multiple data owners within the same domain: An MT system trained with the combined data can deliver significantly improved translation quality, compared to a system trained with the provider's own 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 100000 segments of parallel training data
   – Less than 2M segments of parallel training data (at which point it would be valid to train a System with only the provider's own data)
4. Chosen case: Sybase
5. Experiment Series: Observe BLEU scores using a reserved subset of Sybase's submitted data against systems trained with
   A. General data, as used for www.microsofttranslator.com
   B. Only Microsoft's localization data
   C. Microsoft data + Sybase data
   D. General + Microsoft + TAUS
   E. 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
A	General	General	General
B	Microsoft	Microsoft	Microsoft
C	Microsoft and Sybase	Microsoft and Sybase	Sybase
D	General and Microsoft and TAUS	General Microsoft and TAUS	TAUS
E	General and Microsoft and TAUS	General Microsoft and TAUS Sybase	Sybase

Microsoft Translator

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 – BLEU

Chinese

			Test Set		
System	Size	System Description	General	Microsoft	Sybase
A	8.3M	General domain	14.26	29.74	34.81
B	2.6M	Microsoft	12.32	34.65	29.95
C	2.6M	Microsoft with Sybase	12.16	34.66	30.24
D	11.5M	General and Microsoft and TAUS	15.38	35.80	44.49
E	11.5M	System D with Sybase lambda	12.57	29.51	47.16

German

			Test Set		
System	Size	System Description	General	Microsoft	Sybase
A	4.4M	General Domain	25.19	40.61	34.85
B	7.6M	Microsoft	21.95	52.39	41.55
C	7.6M	Microsoft with Sybase	22.83	52.07	42.07
D	11.1M	General and Microsoft and TAUS	23.86	52.72	48.83
E	11.1M	System D with Sybase lambda	19.44	37.27	50.85

Experiment Results - 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.
- Lamdba training without diversity in the training data has almost no effect (compare B vs. C)
- Lambda training with in-domain diversity has a significant positive effect for the lambda target, and a significant negative effect for everyone else (compare C vs. D)
- 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

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