A Novel Statistical Pre-Processing Model for a Rule-Based Machine Translation System

Yanli Sun, Sharon O'Brien, Minako O'Hagan

School of Applied Language and Intercultural Studies, Dublin City University

Fred Hollowood

Research and Deployment (SES), Symantec Corporation Ireland

Outline

1. Introduction
2. Design
3. Experiments
4. Results

Introduction

- Pre-processing in general:
  - It is the first step in the translation process
  - It prepares the input for effective analysis and transfer for an MT system
  - such as: tokenisation, segmentation, dictionary customisation, Controlled Language (CL) rules, etc.
- Challenges of current pre-processing methods
  - CL (O'Brien, 2003; O'Brien and Roturier, 2007)
    - Rules are manually crafted
    - Hard for writers to implement
  - Source re-construction (Xia and MacCord, 2004; Crego and Marino, 2007; Babych et al., 2009)
    - Focus on SMT systems
    - Sometimes still need to craft rules manually
    - Limited number of rules

Introduction

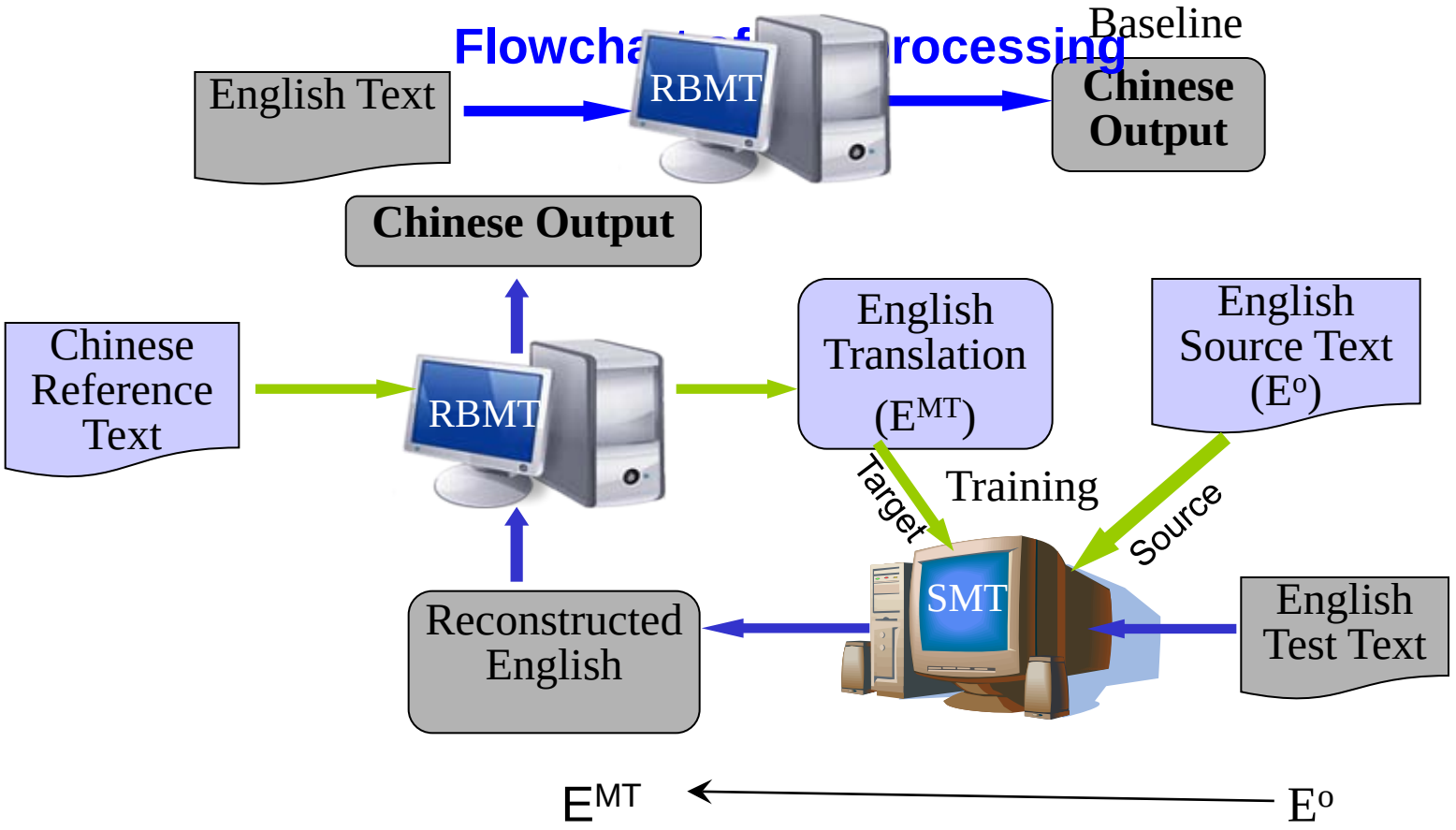
A pre-processing method for RBMT system

- What do we need:
  - Source texts which can be, on the one hand as similar to the Chinese structure as possible, while at the same time, analyzable by the RBMT system
- What do we do:
  - Employ an SMT to transform the English sentence into Chinese friendly (or an RBMT system friendly) structure automatically and comprehensively without human intervention or manually crafted rules

Outline

1. Introduction
2. **Design**
3. Experiments
4. Results

Methodology



symantec.

ENTERPRISE IRELAND

Outline

Experiments

- Experimental set-up
  - MT system: Systran (RBMT) and Moses (SMT)
  - Automatic Evaluation Metrics: GTM, BLEU and TER
  - Training Corpora: Four

The test set and the first corpus belong to a security corpus. The Chinese references were extracted from an in-house TM.

First type Corpus	#Sentence	#English words	#Chinese words
In_Domain	5439	77268	85501
Test Set	944	14839	16100

Second type Corpora	#sentence	#English words	#Chinese words
Mix_Tiny	5439	55846	69410
Mix_Small	9934	106457	119480
Mix_Large	269913	2787175	3382309
Development Set	903	10677	10764

To scale up the project. The whole mix-corpus TM was employed. A development corpus and three other corpora were randomly selected from the TM along with their corresponding references.

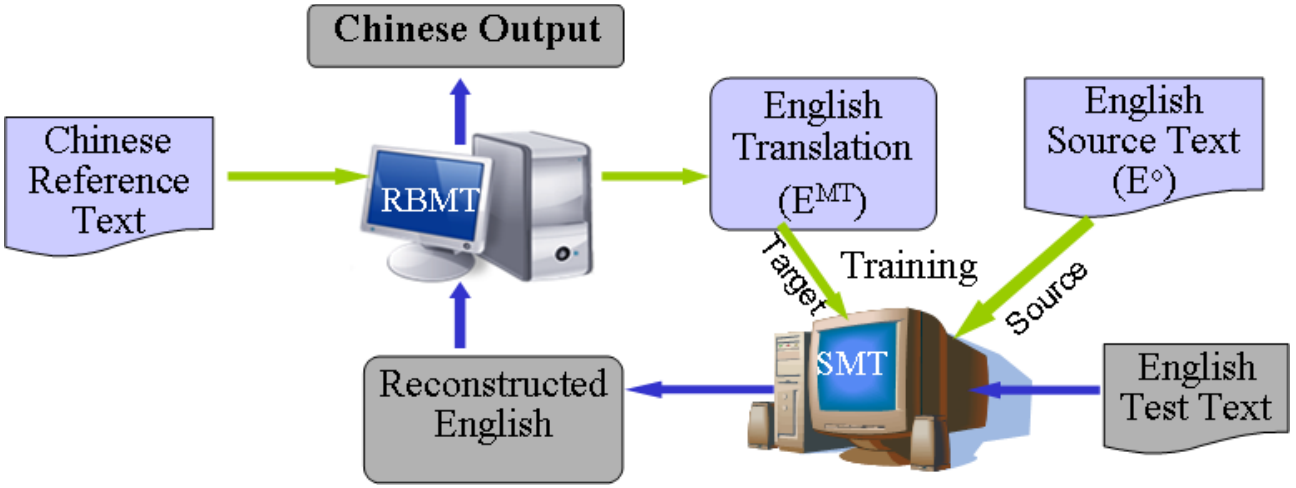
symantec.

ENTERPRISE IRELAND

Outputs

- Baseline

- In_Domain
- Mix_Tiny
- Mix_Small
- Mix_Large

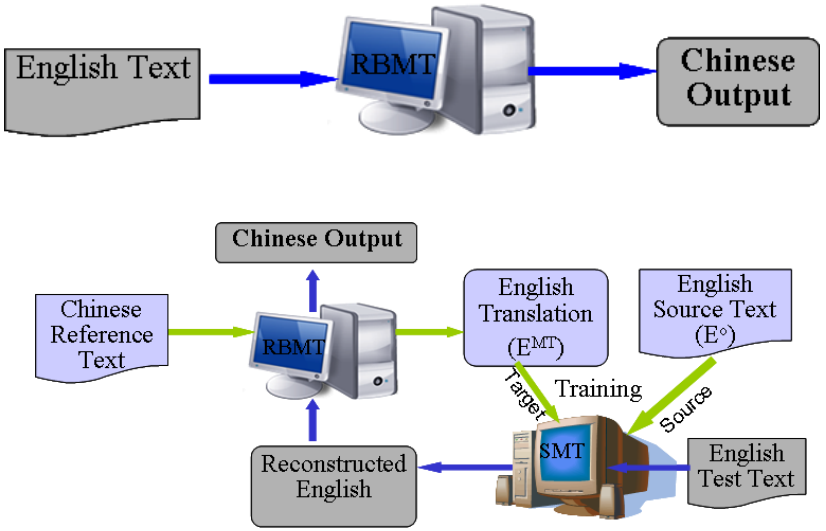


Outline

1. Introduction
2. Design
3. Experiments
4. **Results**

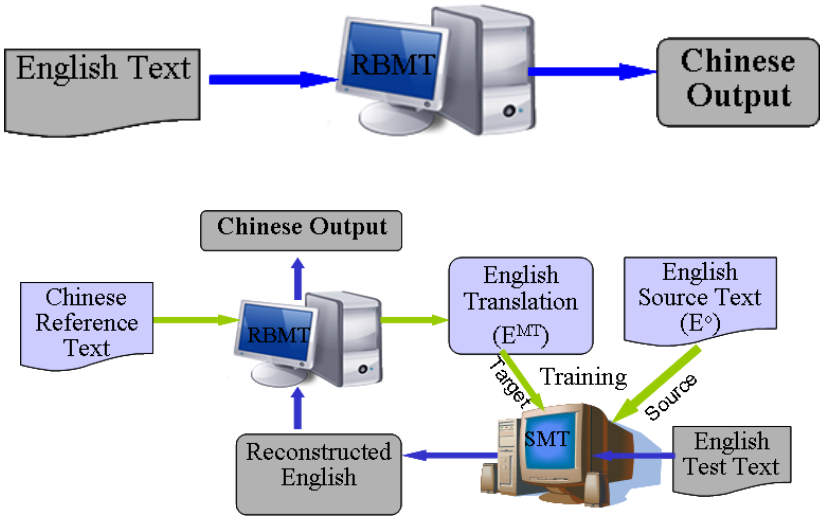
Results - Scores

	GTM	BLEU	TER
Baseline	0.6565	0.2490	0.5249
Mix_Tiny	0.6553	0.2229	0.5499
Mix_Small	0.6567	0.2303	0.5436
Mix_Large	**0.6836**	**0.2746**	**0.5058**
In_Domain	0.6751	0.2646	0.5261



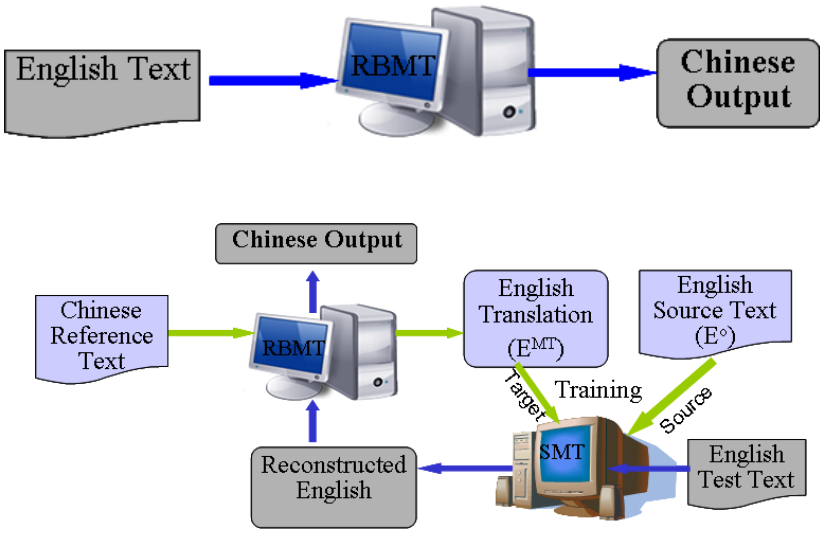
Results - Scores

	GTM	BLEU	TER
Baseline	0.6565	0.2490	0.5249
Mix_Tiny	0.6553	0.2229	0.5499
Mix_Small	0.6567	0.2303	0.5436
Mix_Large	0.6836	0.2746	0.5058
In_Domain	0.6751	0.2646	0.5261



Results - Scores

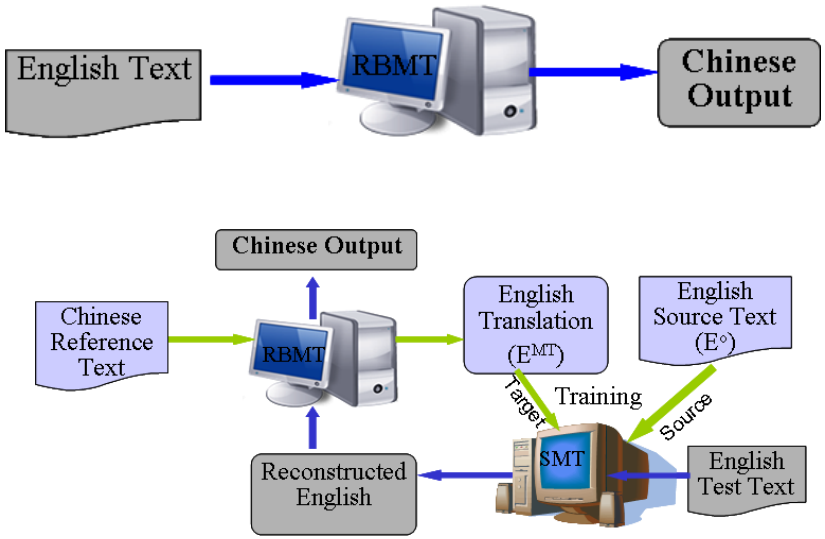
	GTM	BLEU	TER
Baseline	0.6565	0.2490	0.5249
Mix_Tiny	0.6553	0.2229	0.5499
Mix_Small	0.6567	0.2303	0.5436
Mix_Large	0.6836	0.2746	0.5058
In_Domain	0.6751	0.2646	0.5261



Results - Scores

	GTM	BLEU	TER
Baseline	0.6565	0.2490	0.5249
Mix_Tiny	0.6553	0.2229	0.5499
Mix_Small	0.6567	0.2303	0.5436
Mix_Large	**0.6836****	**0.2746****	**0.5058****
In_Domain	**0.6751****	0.2646*	0.5261

**p<0.001; *p<0.1



- The biggest corpus produced the best translation according to the scores;
- In-domain corpus is better than mix-domain corpus unless the mix-domain corpus is greatly larger than the in-domain corpus

Results – Translation Comparison

- Baseline vs. Mix_Large (the biggest corpus), an example:

Baseline (Without Pre-processing)

English Source	About the processes that proactive threat scans detect							
Ref / Baseline	关于	主动型	威胁	扫描	所	检测	的	进程
关于	■							
主动型		■						
威胁			■					
扫描				■				
的							■	
进程								■
请								
检测						■		
Gloss	About the processes of proactive threat scans please detect							

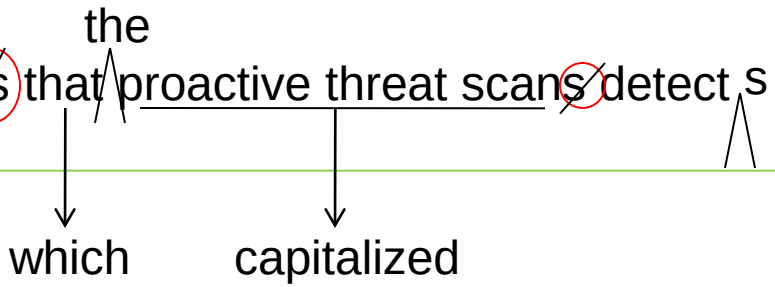
Mix_Large (With Pre-processing)

English Source	About the processes that proactive threat scans detect							
Ref / Mix_Large	关于	主动型	威胁	扫描	所	检测	的	进程
关于	■							
主动型		■						
威胁			■					
扫描				■				
检测						■		
的							■	
进程								■
Gloss	About the process that proactive threat scans detect							

Source English VS. Mix_Large English

Source sentence:

* About the processes that proactive threat scans detect s



Pre-processed sentence:

* About process which the Proactive Threat Scan detects

Source English VS. Mix_Large English

The most frequent changes made by the pre-processing model

Category (# occurred)	Example	Frequency	
Insertion (1158)	the		248
	will		46
	,		41
	”		39
	to		36
Deletion (992)	the		102
	of		85
	a		65
	that		59
	you		49
Substitution (5307)	a	the	166
	can	may	150
	computer	machine	64
	that	which	58
	click	clicks	49

symantec.

ENTERPRISE IRELAND

Source English VS. Mix_Large English

- 99.8% (942 out of 944) of sentences were modified
  - All the modified English sentences could be divided into three groups:

	Percentage
Correct grammar and clear meaning	25.64
Minor error and clear meaning	25.74
Incorrect grammar and fuzzy meaning	48.31

Source English VS. Mix_Large English

- 99.8% (942 out of 944) of sentences were modified
  - All the modified English sentences could be divided into three groups:

	Percentage	
Correct grammar and clear meaning	25.64	20% sentences in this group changed meaning
Minor error and clear meaning	25.74	
Incorrect grammar and fuzzy meaning	48.31	

Source English VS. Mix_Large English

- 99.8% (942 out of 944) of sentences were modified
  - All the modified English sentences could be divided into three groups:

	Percentage	
Correct grammar and clear meaning	25.64	20% sentences in this group changed meaning
Minor error and clear meaning	25.74	
Incorrect grammar and fuzzy meaning	48.31	

symantec.

ENTERPRISE IRELAND

Source English VS. Mix_Large English

- 99.8% (942 out of 944) of sentences were modified
  - All the modified English sentences could be divided into three groups:

	Percentage	
Correct grammar and clear meaning	25.64	20% sentences in this group changed meaning
Minor error and clear meaning	25.74	
Incorrect grammar and fuzzy meaning	48.31	

Question: Which group of sentences produce most of the improvements?

symantec.

ENTERPRISE IRELAND

Source English VS. Mix_Large English

Pilot Human Evaluation

- 3 students on 100 sentences
- Compare the translations of sentences within each group to the baseline translation
- Calculate the improvements and degradations ratios

	Improvements / Degradations
Correct grammar and clear meaning	0.6970
Minor error and clear meaning	0.7941
Incorrect grammar and fuzzy meaning	0.5517

No significant difference between baseline and pre-processed translation

Conclusions

- Automatic pre-processing for RBMT system can be conducted using an SMT system
- An increase of about 10% automatic scores were reported
- In-domain corpus performs better than the same sized or slightly bigger mix-domain corpora but not as good as huge corpus
- Degradations were generated by the SMT system

Next steps

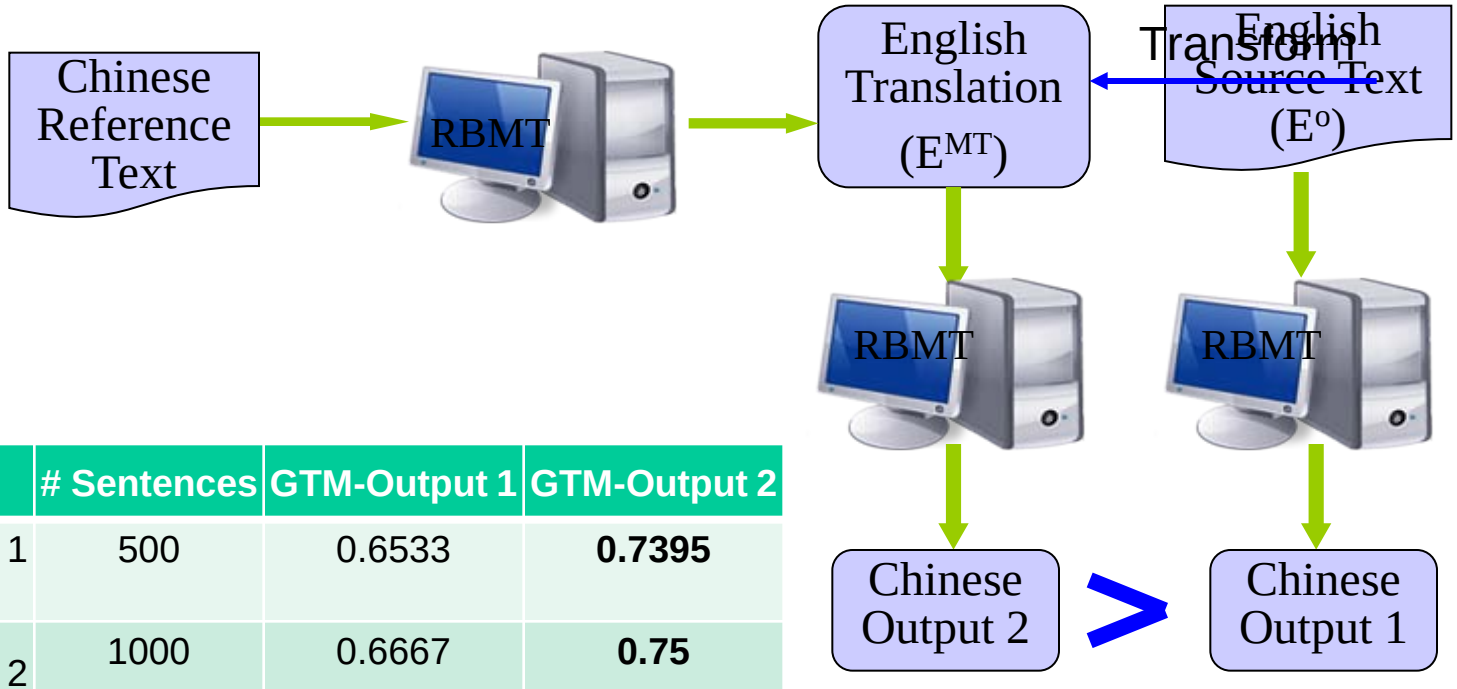
- How to improve the performance of the model
  - Regulate the SMT system to only pre-process sentences meeting certain criteria?
  - Regulate the RBMT system to translate only grammatical or ungrammatical sentences?
  - Use cleaned and balanced corpora?

Questions?

Suggestions?

Thanks!

Rationale



	# Sentences	GTM-Output 1	GTM-Output 2
Sample 1	500	0.6533	**0.7395**
Sample 2	1000	0.6667	**0.75**

Result – Linguistic Analysis

- 99.8% (942 out of 944) of sentences were modified
  - All the modified English sentences could be divided into three groups:

1. Correct English grammar and clear meaning (25.64%)
   - Keep the original meaning (80%), e.g.

     Source: You know that the process is safe to run in your environment.

     Pre-processed: You know that the procedure is safe to run in your conditions.

     - Changed meaning (20%), e.g.

     Source: You configure Auto-Protect settings as part of an Antivirus and Antispyware Policy.

     Pre-processed: You configure the auto-Protect settings in antivirus and antispyware tactic.

Result – Linguistic Analysis

2. Minor English grammar error and understandable meaning (25.74%):
   - All sentences keep the original meaning

     Source: Auto-Protect then scans the files if you request them from the remote computer again.

     Pre-processed: The auto-protect will scan this file, if your request them from the remote machine once more.

3. Incorrect English grammar and hard to comprehend (48.31%)

   Source: After Policy name, type the name of the policy (it shows New Host Integrity Policy by default).

   Pre-processed: After policy name, the type policy name (displays “new Host Integrity Policy” default).