

Self-training for Machine Translation

Nicola Ueffing

**Interactive Language Technologies Group
NRC Canada**





- **Introduction**
- **Self-training for SMT**
- **Extensions and Modifications**
- **Experimental results**
- **Conclusion**

Existing statistical machine translation systems benefit from

- **bilingual parallel or comparable corpora in the source and target language**
- **monolingual corpora in the target language**

They do *not* benefit from the availability of monolingual corpora in the source language

Bilingual parallel data

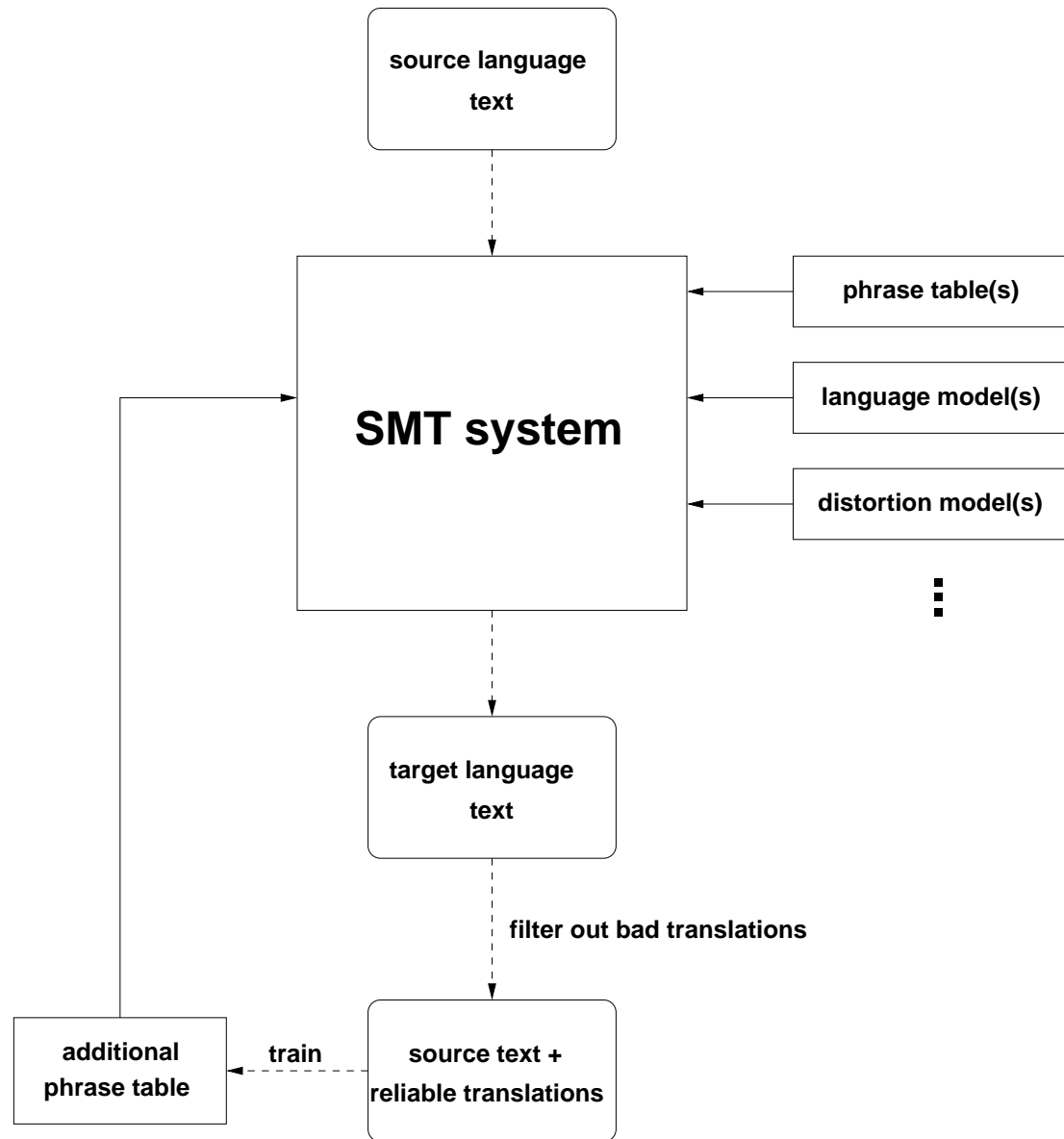
- **often only limited amount available**
- **creation expensive**

Here: explore monolingual source-language corpora to improve translation quality

- **adapt to new domain, topic or style**
- **overcome training/testing data mismatch, e.g. text/speech**

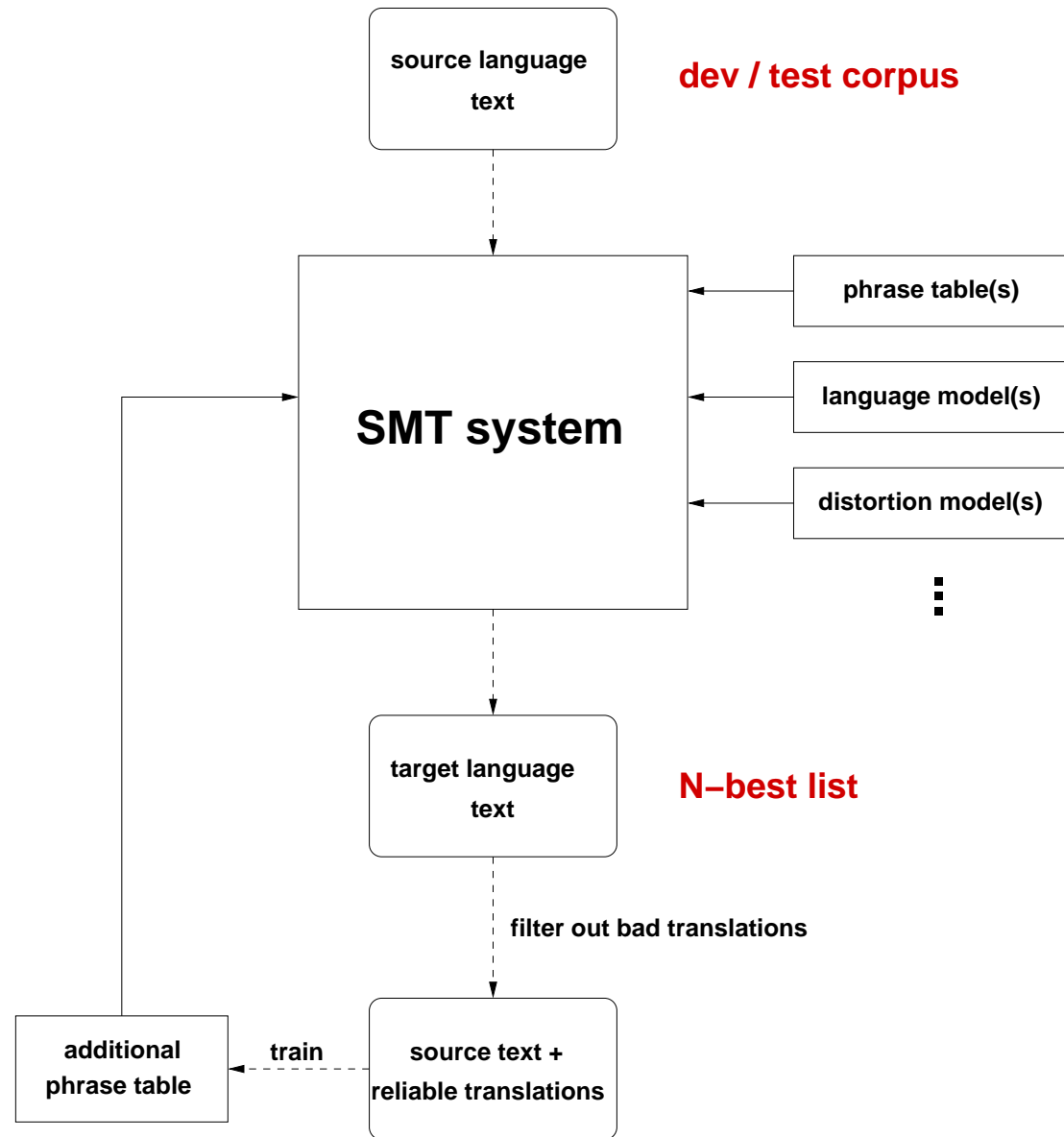


Self-training (1)



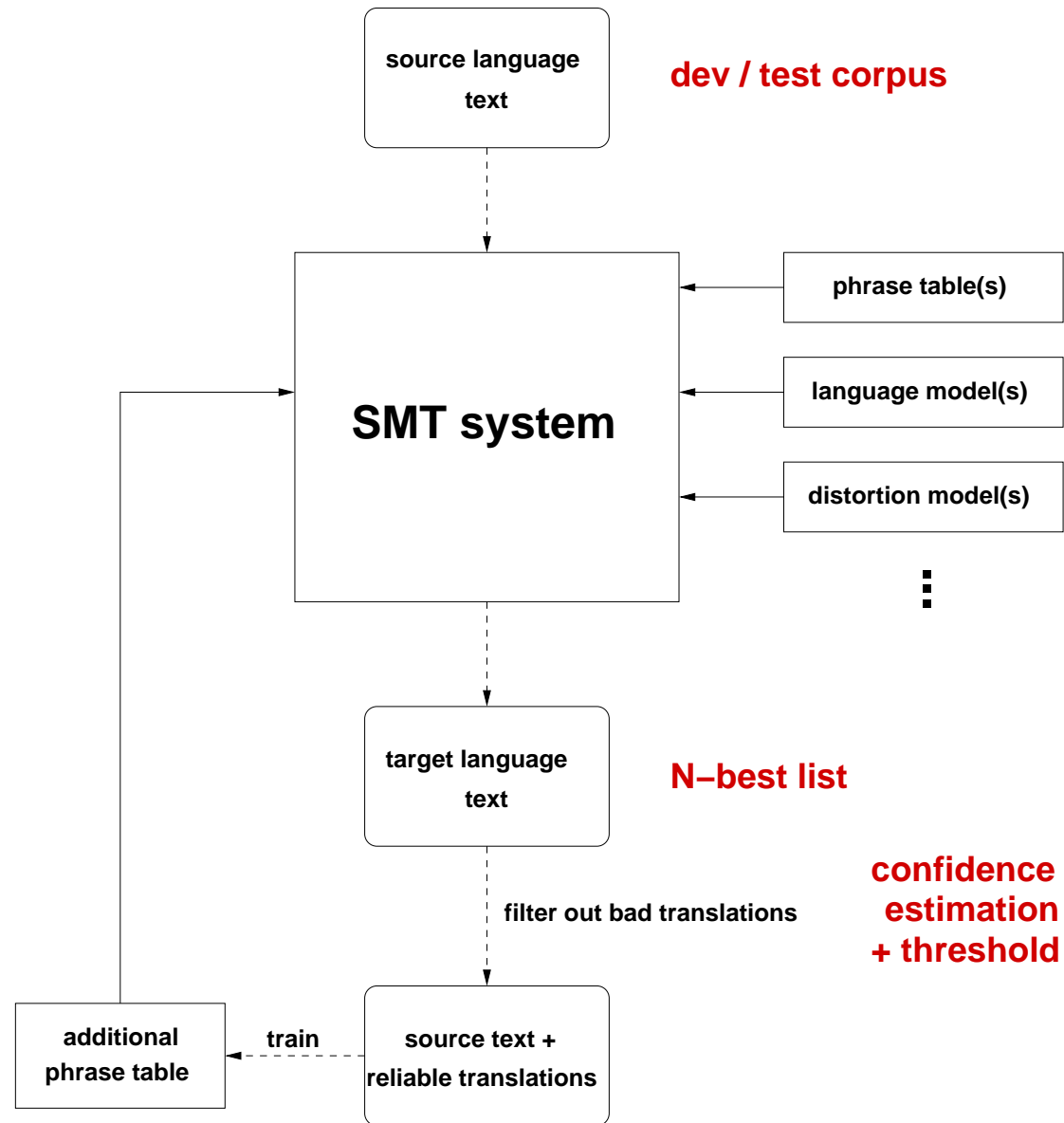


Self-training (1)





Self-training (1)





Why does retraining the SMT system on its own output work?

- Reinforce parts of the phrase translation model which are relevant for test corpus, obtain more focused probability distribution
- Compose new phrases, example:

original parallel corpus

'A B', 'C D E'

additional SL data

'A B C D E'

possible new phrases

'A B C', 'B C D E', 'A B C D E', ...

Why does retraining the SMT system on its own output work?

- Reinforce parts of the phrase translation model which are relevant for test corpus, obtain more focused probability distribution
- Compose new phrases, example:

original parallel corpus	additional SL data	possible new phrases
'A B', 'C D E'	'A B C D E'	'A B C', 'B C D E', 'A B C D E', ...

Limitations of the approach:

- No learning of translations of *unknown* source-language words occurring in the new data
- Only learning of *compositional* phrases; system will not learn translation of idioms:

"it is raining"	+	"cats and dogs"	→	"it is raining cats and dogs"
"es regnet"	+	"Katzen und Hunde"	↯	"es regnet in Strömen"
"il pleut"	+	"des chats et des chiens"	↯	"il pleut à boire debout"



Iterative Procedure

- 1. Translate test corpus using baseline system**
- 2. Filter out bad translations and learn phrases**
- 3. Train adapted system**
- 4. Repeat steps 1 to 3 using adapted system**



Impact of confidence estimation

How important is it to filter out bad translations? Vary threshold:

- accept *all* translations
- adjust confidence threshold and keep approx. 1/6, 1/3, 1/2, 2/3 of machine translations



***N*-best lists of machine translations**

- **allow for alternative translations of a source sentence**
- **keep up to n confident hypotheses from *N*-best list**
- **study different values of n**



Overview

- **experimental setting, evaluation**
- **SMT system**
- **experimental results**



Setup and evaluation:

- **Chinese → English translation**
- **training conditions: NIST 2006 eval, large data track**
- **testing: 2004 and 2006 eval corpora, 3 and 4 different genres, partially not covered by training data (broadcast conversations, speeches, ...)**
- **learn separate phrase table for each genre**
- **evaluate with BLEU-4, mWER, mPER, using 4 / 1 references**
- **95%-confidence intervals, using bootstrap resampling**



PORTAGE phrase-based statistical translation system

Decoder models:

- **several (smoothed) phrase table(s), translation direction $p(s_1^J | t_1^I)$**
- **several 4-gram language model(s), trained with SRILM toolkit**
- **distortion model, penalty based on number of skipped source words**
- **word penalty**

Combine models log-linearly, optimize weights w.r.t. BLEU score using Och's algorithm

Search: dynamic-programming beam-search algorithm

Additional rescoring models:

- **two different IBM-1 features, each in both translation directions**
- **posterior probabilities for words, phrases, n -grams, and sentence length: calculated over the N -best list, using the sentence probabilities assigned by the baseline system**

Self-training

Translation quality on the NIST Chinese–English task.

corpus	system	BLEU[%]	mWER[%]	mPER[%]
eval-04	baseline	31.8 ±0.7	66.8 ±0.7	41.5 ±0.5
	adapted	33.5	65.3	40.8
eval-06	GALE baseline	12.7 ±0.5	75.8 ±0.6	54.6 ±0.6
	GALE adapted	13.6	73.4	53.2
NIST	baseline	27.9 ±0.7	67.2 ±0.6	44.0 ±0.5
	adapted	29.3	65.6	43.2

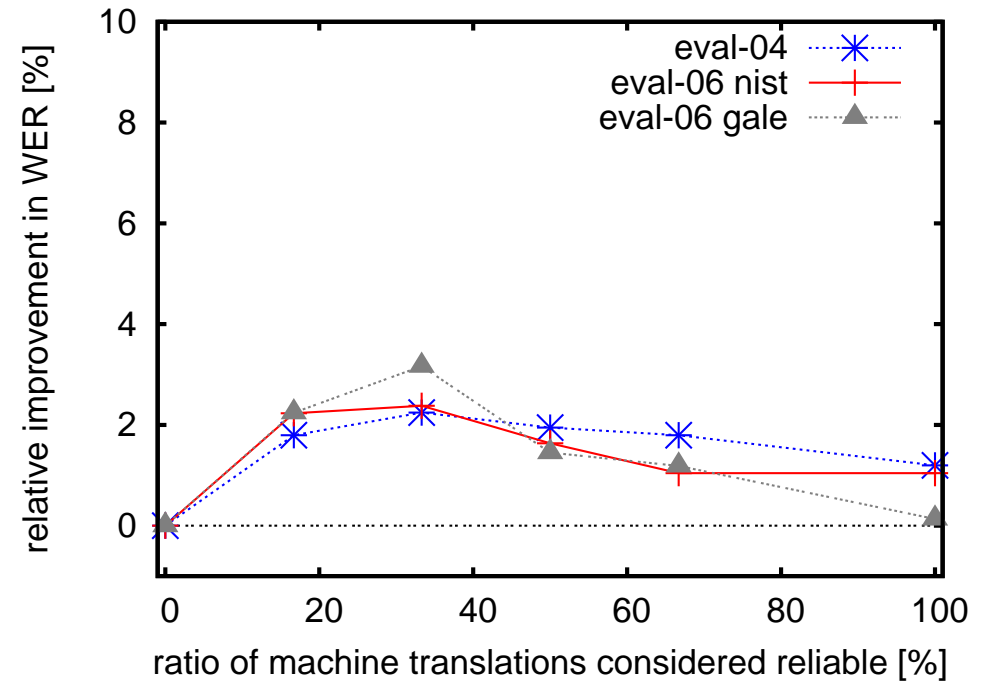
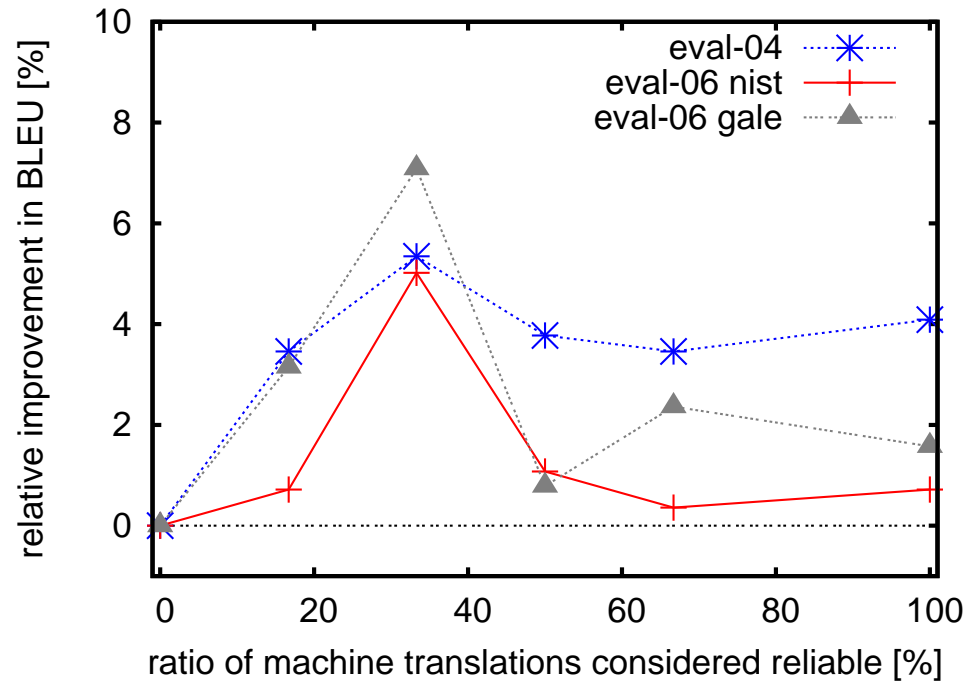


Iterations

Translation quality on the NIST Chinese–English task.

corpus	system	BLEU[%]	mWER[%]	mPER[%]	
eval-04	baseline	31.8 ±0.7	66.8 ±0.7	41.5 ±0.5	
	iter 1	33.5	65.3	40.8	
	iter 2	32.9	65.0	40.9	
eval-06	GALE	baseline	12.7 ±0.5	75.8 ±0.6	54.6 ±0.6
		iter 1	13.6	73.4	53.2
		iter 2	13.0	73.3	54.1
	NIST	baseline	27.9 ±0.7	67.2 ±0.6	44.0 ±0.5
		iter 1	29.3	65.6	43.2
		iter 2	28.1	65.0	43.7

Threshold





***N*-best**

Translation quality in terms of BLEU[%] on the NIST Chinese–English task.

corpus	baseline	$N = 1$	2	5	10	100	
eval-04	31.8	33.5	32.8	33.1	33.5	33.2	
eval-06	GALE	12.7	13.6	13.1	13.3	12.5	13.1
	NIST	27.9	29.3	27.9	28.2	28.1	28.6



- **explore monolingual source-language data to improve an existing MT system:**
 - **translate data using MT system**
 - **automatically identify reliable translations**
 - **learn new models on these**
- **translation quality improves through self-training**
- **filtering based on confidence of translations is important**
- **iterative procedure does not improve translation quality**
- **consider only single-best of the translations in self-training**



- **Self-training: Ueffing [IWSLT, 2006]**
- **PORTAGE: Johnson et. al. [NAACL SMT Workshop, 2006]**
- **Minimum error rate training: Och [ACL 2003]**
- **Phrase extraction: Koehn et. al. [HLT/NAACL, 2003]**
- **Phrase table smoothing: Foster et. al. [EMNLP, 2006]**
- **Confidence measures: Blatz et. al. [JHU workshop final report, 2003], Ueffing and Ney [to appear in CompLing 33-01, 2007]**



END