Using Monolingual Source-Language Data to Improve MT Performance

Nicola Ueffing

Interactive Language Technologies Group
NRC Canada

National Research Council Canada
Conseil national de recherches Canada

Canada

Overview

- **Introduction**
- **Review of statistical machine translation (SMT)**
- **Use of monolingual data**
- **Experimental results**
- **Conclusion**

Motivation

Existing statistical machine translation systems benefit from

- **bilingual parallel or comparable corpora in the source and target language**
- **monolingual corpora in the target language**
- **multilingual parallel corpora (multi-source translation, co-training)**

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**

Statistical machine translation

Goal: minimization of decision errors for translation
generate target sentence with the largest posterior probability

$$\begin{aligned} s_1^J \rightarrow (\hat{I}, \hat{t}_1^{\hat{I}}) &= \underset{I, t_1^I}{\operatorname{argmax}} \{pr(t_1^I \mid s_1^J)\} \\ &= \underset{I, t_1^I}{\operatorname{argmax}} \{pr(t_1^I) \cdot pr(s_1^J \mid t_1^I)\} \end{aligned}$$

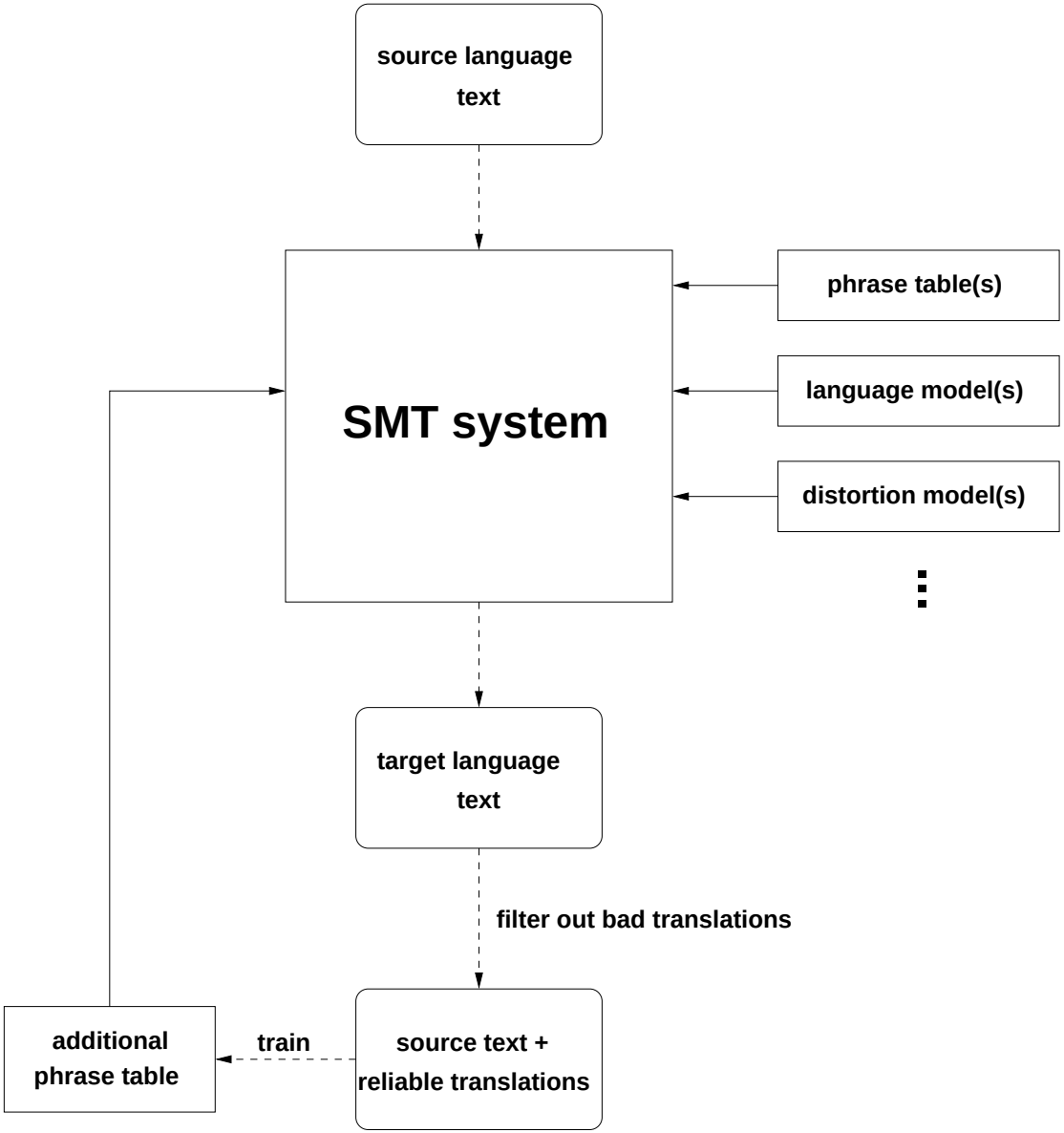
State-of-the-art systems combine many different knowledge sources:

- **score based on source and target sentence:** $g_k(s_1^J, t_1^I)$
- **score based on target sentence only:** $h_l(t_1^I)$

Resulting search criterion:

$$\underset{I, t_1^I}{\operatorname{argmax}} \left\{ p(s_1^J | t_1^I)^{\alpha_0} \cdot p(t_1^I)^{\beta_0} \cdot \prod_{k=1}^{K} g_k^{\alpha_k}(s_1^J, t_1^I) \cdot \prod_{l=1}^{L} h_l^{\beta_l}(t_1^I) \right\}$$

Use of monolingual data (1)



Use of monolingual data (2)

Procedure:

1. **Translate new source text using existing MT system, generate 5k-best lists.**
   Here: dev/test corpus $\Rightarrow$ adapt to its topic, domain, style, . . .
   small and very specific phrase table

Use of monolingual data (2)

Procedure:

1. **Translate new source text using existing MT system, generate 5k-best lists.**
   Here: dev/test corpus $\Rightarrow$ adapt to its topic, domain, style, . . .
   small and very specific phrase table
2. **Estimate confidence $c(t_1^I)$ of each single-best translation: log-linear combination of different posterior probabilities calculated over N-best list and LM score.**

Use of monolingual data (2)

Procedure:

1. **Translate new source text using existing MT system, generate 5k-best lists.**
   Here: dev/test corpus $\Rightarrow$ adapt to its topic, domain, style, . . .
   small and very specific phrase table
2. **Estimate confidence $c(t_1^I)$ of each single-best translation: log-linear combination of different posterior probabilities calculated over N-best list and LM score.**
3. **Identify reliable translations based on confidence scores**
   - **compare $c(t_1^I)$ to threshold τ**
     $c(t_1^I) \geq \tau$: consider t_1^I reliable and keep it
     $c(t_1^I) < \tau$: discard t_1^I and leave s_1^J untranslated
   - **build new bilingual corpus from reliable translations and their sources**

Use of monolingual data (3)

4. Train new model $g_k(s_1^J, t_1^I)$ on reliable translations and use this as additional feature function in the existing system. Here: new phrase table
   - using IBM models and Koehn's "diag-and" method
   - maximal phrase length of 4 words (original tables: 5 and 8)
   - smooth tables using Kneser-Ney and lexical Zens-Ney
5. Adapted SMT system
   - all original models
   - one new phrase table (specific to each test corpus)
   - new optimization of weights

Use of monolingual data (4)

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"	$\rightarrow$	"it is raining cats and dogs"
"es regnet"	+	"Katzen und Hunde"	$\not\rightarrow$	"es regnet in Strömen"
"il pleut"	+	"des chats et des chiens"	$\not\rightarrow$	"il pleut à boire debout"

Experimental results

Overview

- experimental setting, evaluation
- SMT system
- experimental results
- translation examples

Experimental setting

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**
- **pairwise comparison using bootstrap resampling**

Experimental setting: SMT system

PORTAGE phrase-based statistical translation system

Decoder models:

- **several phrase table(s), translation direction $p(s_1^J \mid t_1^I)$**
- **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**

Experimental results (1)

Translation quality on the NIST Chinese–English task

corpus		system	BLEU[%]	mWER[%]	mPER[%]
eval-04^{+}		baseline	31.8	66.8	41.5
		adapted	32.6	65.3	40.8
eval-06	GALE*	baseline	12.7	75.8	54.6
		adapted	13.3	73.6	53.4
	NIST^{+}	baseline	27.9	67.2	44.0
		adapted	28.4	65.9	43.4

$^{+}$ 4 references * 1 reference

Experimental results (2)

Translation quality, separate evaluation for each genre

corpus			system	BLEU[%]	mWER[%]	mPER[%]
eval-04+		editorials	baseline	30.7	67.0	42.3
			adapted	31.8	65.7	41.8
		newswire	baseline	30.0	69.1	42.7
			adapted	31.1	67.1	41.8
		speeches	baseline	36.1	62.5	38.6
			adapted	36.1	61.7	38.3
eval-06	GALE*	broadcast conversations	baseline	10.8	78.7	59.2
			adapted	11.9	75.6	56.9
		broadcast news	baseline	12.3	76.7	54.0
			adapted	12.8	75.0	53.2
		newsgroups	baseline	11.8	73.6	55.1
			adapted	12.2	71.7	54.3
		newswire	baseline	16.2	72.9	49.0
			adapted	16.5	70.8	48.0
eval-06	NIST+	broadcast news	baseline	29.5	66.8	44.3
			adapted	30.5	65.5	42.7
		newsgroups	baseline	22.9	68.2	48.8
			adapted	22.5	66.8	48.2
		newswire	baseline	29.1	67.0	41.6
			adapted	29.9	65.6	41.4

+ 4 references * 1 reference

Experimental results (3)

Summary

- system adapts well to new topics and domains
- largest gains on broadcast conversations (not contained in training data)
- 28% – 48% of the adaptive phrases are new (not in original phrase tables), but system uses mainly “old” phrase pairs
- 40% of the adaptive phrases are used in generation of rescored-best translations

Experimental results: examples

Translation examples from the 2006 GALE corpus

baseline	the report said that the united states is a potential problem, the practice of china's foreign policy is likely to weaken us influence.
adapted	the report said that this is a potential problem in the united states, china is likely to weaken the impact of american foreign policy.
reference	the report said that this is a potential problem for america. china's course of action could possibly weaken the influence of american foreign policy.
baseline	the capitalist system, because it is immoral to criticize china for years, capitalism, so it didn't have a set of moral values .
adapted	capitalism has a set of moral values, because china has denounced capitalism, so it does not have a set of moral.
reference	capitalism, its set of morals, because china has criticized capitalism for many years, this set of morals is no longer there.
baseline	what we advocate his name
adapted	we advocate him.
reference	we advocate him.
baseline	the fact that this is.
adapted	this is the point.
reference	that is actually the point.
baseline	"we should really be male nominees .. ....
adapted	he should be nominated male, really.
reference	he should be nominated as the best actor, really.

Conclusion

- 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
- improvements in settings where testing data does not match training data well
- lots of ongoing and future work. . .

Literature

- **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

Corpus statistics

Chinese–English Corpora

corpus	use	# sentences	domains
non-UN	phrase table + LM	3,164,180	news, magazines, laws, Hansards
UN	phrase table + LM	4,979,345	UN Bulletin
English Gigaword	LM	11,681,852	news
multi-p3	dev1	935	news
multi-p4	dev2	919	news
eval-04	test	1,788	newswire (NW), editorials (ED), political speeches (SP)
eval-06	test	3,940	broadcast conversations (BC), broadcast news (BN), newsgroups (NG), newswire (NW)

Experimental results (1)

Statistics of the phrase tables trained on the different genres of the test corpora.

corpus		# sentences	# reliable translations	phrase table size	# adapted phrases used	# new phrases	# new phrases used
eval-04	ED	449	101	1,981	707	679	23
	NW	901	187	3,591	1,314	1,359	47
	SP	438	113	2,321	815	657	25
eval-06	BC	979	477	2,155	759	1,058	90
	BN	1,083	274	4,027	1,479	1,645	86
	NG	898	226	2,905	1,077	1,259	88
	NW	980	172	2,804	1,115	1,058	41