Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation

Holger Schwenk

LIUM, University of Le Mans, France

Holger.Schwenk@lium.univ-lemans.fr

October 20, 2008

Introduction

Training of an SMT system

- Only bilingual sentence-aligned texts ("*bitexts*") and large monolingual texts are needed
- An SMT system can be developed without the need of any language-specific expertise
- Monolingual data is usually available in large amounts
- But aligned bilingual texts are a sparse resource for many language pairs (too small, out-of-domain, ...)

Introduction

How to resolve the problem of insufficient bitexts ?

- Pay people to produce more bitexts
- Integration of high quality dictionaries
- Try to take better advantage of limited data (factored translation model, ...)
- Get more bitexts from the Internet:
  - Most of the found bilingual texts are not direct translations of each other that can be easily aligned

  $=$ *comparable corpora* (Wikipedia, international news agencies, ...)
  - How to exploit comparable corpora ?

  $\Rightarrow$ Try to align some of the sentences [Munteanu and Marcu CL'05, Resnik and Smith CL'03, ...]

Unsupervised Training of an SMT system

Our approach

- Build a baseline SMT system
  (using a limited amount of bitext or out-of-domain)
- Use this system to translate large amounts of texts in the source language
- Build a new SMT system with these translations together with the source as additional bitexts
- We don't need a comparable corpus,
  just texts in the source language

Variants

- Add related translations to the target LM

$\Rightarrow$ Light supervision using a comparable corpus

- How good should be the initial SMT system ?

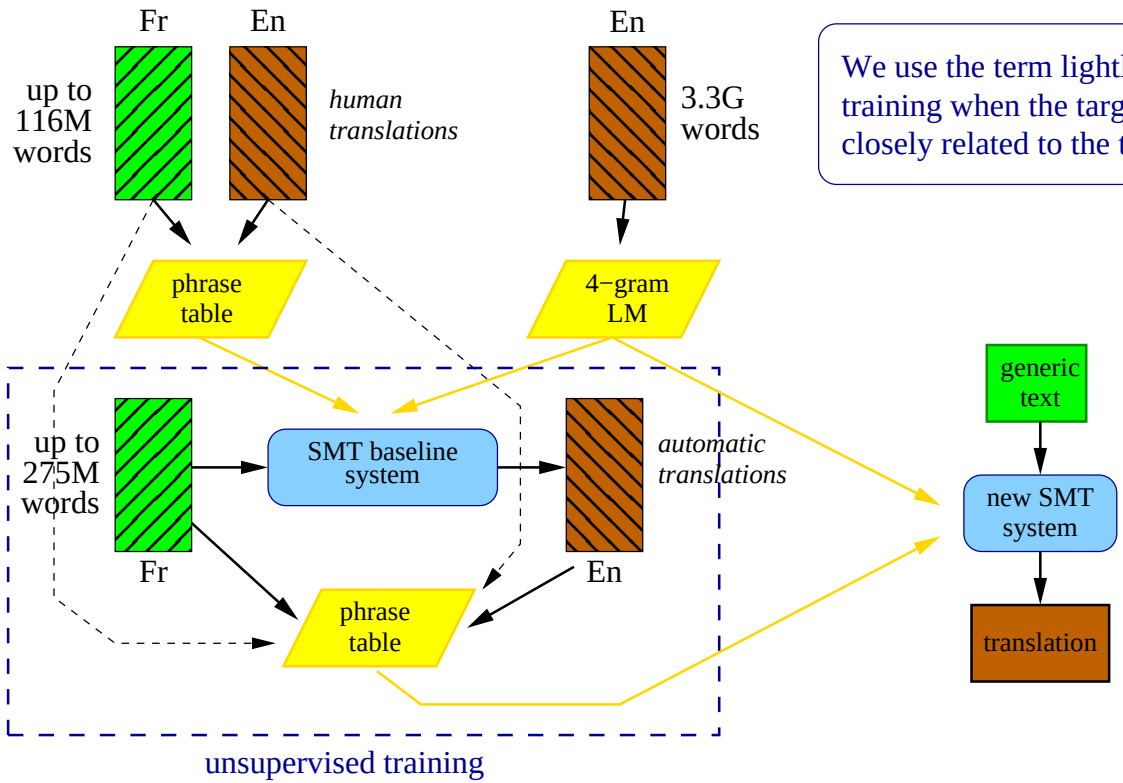
Unsupervised Training of an SMT system

Setup

- Try to build a generic news translation system (French $\longrightarrow$ English)
- Lightly-unsupervised training on LDC's Gigaword corpora:

	#Words	
corpus	French	English
AFP 199x	236M	132M
AFP 200x	337M	326M
Europarl	*40M*	*36M*

Principle of Lightly-Supervised Training



We use the term lightly−supervised training when the target LM data is closely related to the text to be translated

Related Work

Unsupervised training for domain adaptation

- Several papers by Ueffing et al. [IWSLT'06, ACL'07]
  - translate the test data
  - compute confidence scores and filter the outputs
  - adapt the system (build small additional phrase table, ...)
- Work by Chen et al [ACL'08]

- Produce bitexts with a Rule-based system [Hu et al, EMNLP'07]

Baseline System (1)

- Based on the Moses toolkit

Dev and Test data

- Newstest2008 from WMT'08 evaluation
- Consists of news texts (politics, health, financial, society, music, ...) collected from the Internet
- Split randomly into Dev and Test set
  (about 1000 lines, 22k words each)
- Many spelling errors in the French translations were automatically corrected

Baseline System (2)

Available Bitexts

- News-commentary corpus (1.6M words),
- Europarl corpus (40.1M words),
- Canadian Hansard corpus (72.4M words).
- Bilingual dictionary from SYSTRAN (512k words).

Available LM data

- English part of bitexts
- UN data
- Full English Gigaword

Example Translations (1)

French source text:

- *La paix exige une direction palestinienne nouvelle et différente, afin que puisse naître un Etat palestinien. J'appelle le peuple palestinien à élire de nouveaux dirigeants, des dirigeants qui ne soient pas compromis avec le terrorisme.*

Automatic translations:

- *The peace requires a new and different Palestinian leadership, so that we can create a Palestinian state. I call on the Palestinian people to elect new leaders, leaders not compromised by terrorism.*

Example Translations (2)

French source text:

- *M. Arafat, qui s'est juré de faire de l'année 2000 celle de la proclamation d'un Etat palestinien, a mis un point d'honneur à recevoir les six chefs d'Etat présents.*

Automatic translations:

- *Mr. Arafat, who has vowed to make the year 2000 the proclamation of a Palestinian state, has made a point of honour to receive the six heads of state present.*

Example Translations (3)

French source text:

- *Trois heures après, c'était au tour de la Colombie britannique et de Vancouver de célebrer l'arrivée de l'an nouveau.*

Automatic translations:

- *Three hours later, it was the turn of the British Columbia and Vancouver célebrer the arrival of the new year.*

Performance of Baseline Systems

Bitexts	Dict.	Words	Dev	Test
Big SMT system				
News-commentary + Eparl + Hansard	+	116M	22.69	22.17

Performance of Baseline Systems

Bitexts	Dict.	Words	Dev	Test
Big SMT system				
News-commentary + Eparl + Hansard	+	116M	22.69	22.17
Small SMT system				
News-commentary	+	2.4M	20.44	20.18
News-commentary	-	1.6M	19.41	19.53
News-commentary + Eparl	+	43.3M	22.27	22.35

The big SMT system was ranked best in the 2008 WMT evaluation

Some Comments on the Dictionary

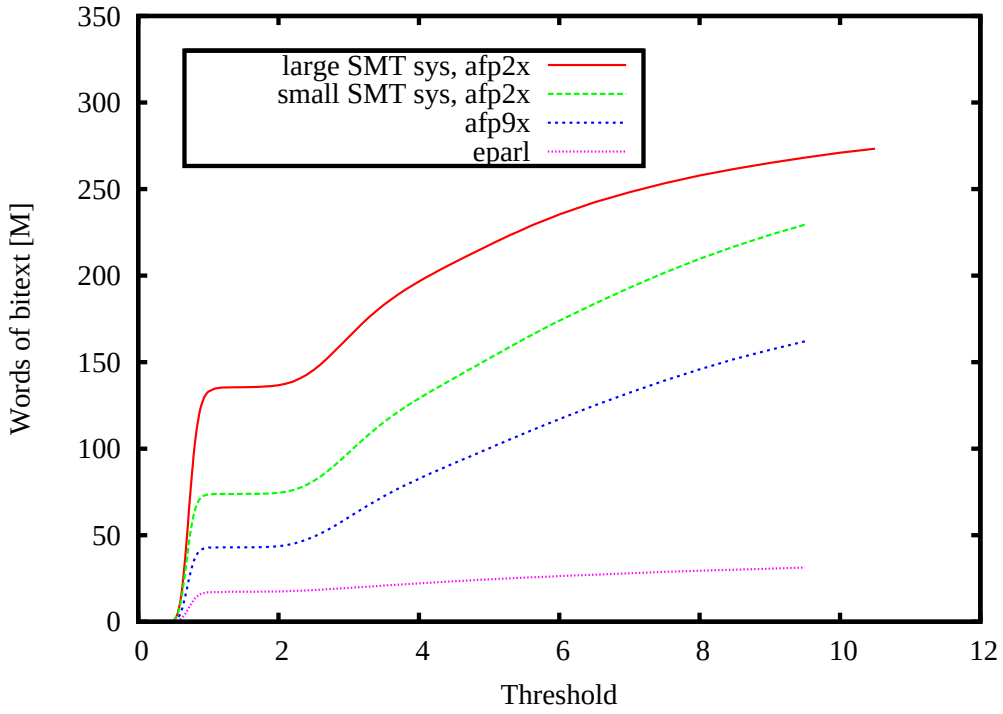
- Provides nouns in singular and plural
- Verbs and in all tenses, ...
- These dictionary entries were directly added to the bitexts
  - Can potentially improve the other alignments
  - Words appearing only in the dictionary have bad translation probabilities

$\Rightarrow$ We hope to improve these probabilities by lightly-supervised training

- Many of the dictionary entries are likely to appear in the large monolingual texts

Filtering the Automatic Translations

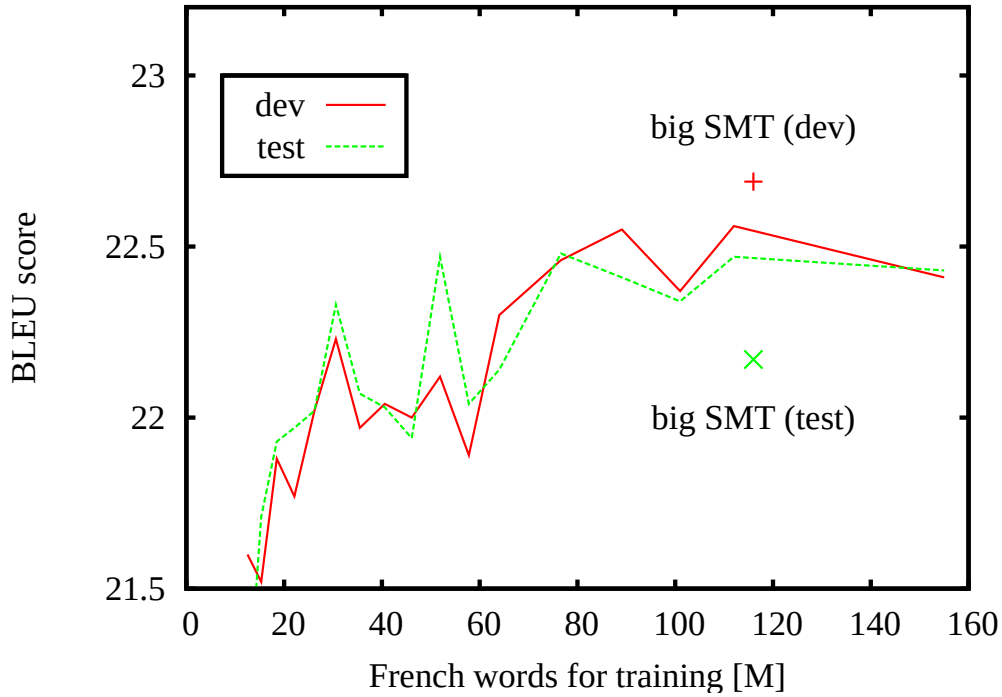
- Try to discard the bad translations
- Some are tables or enumerations of names, places, ...
- We just used the normalized sentence likelihood



$\Rightarrow$ Use up to 150M words of automatic translations

Using the large SMT Baseline System

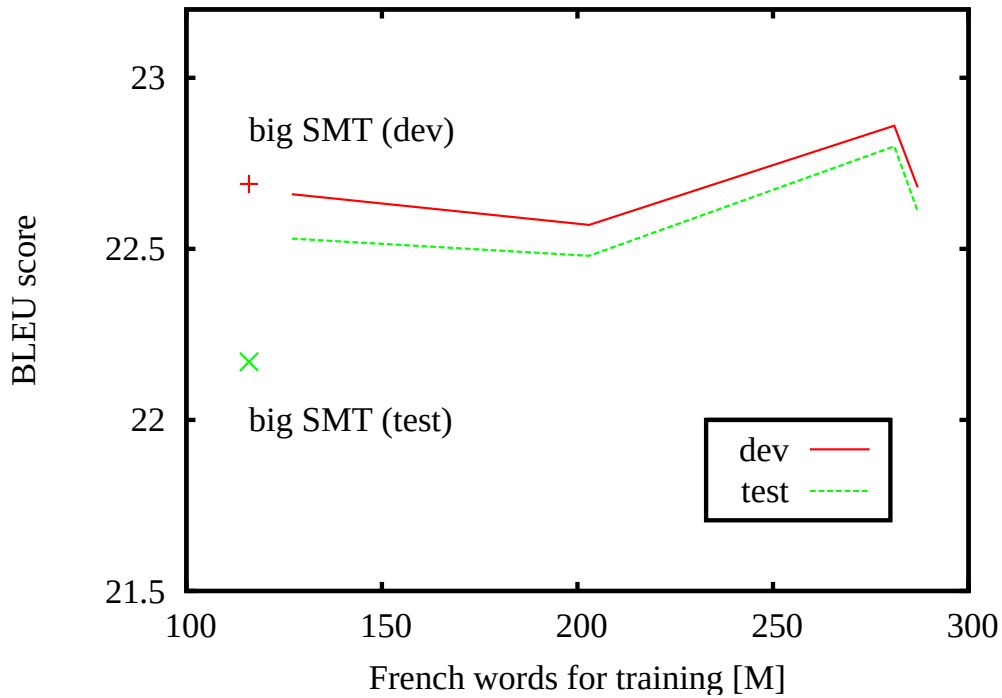
- Build SMT system with automatic translations only



- Better than the baseline when using more than 70M words
- Seems to generalize better
- Improved translation probabilities for dictionary words ?

Using the large SMT Baseline System

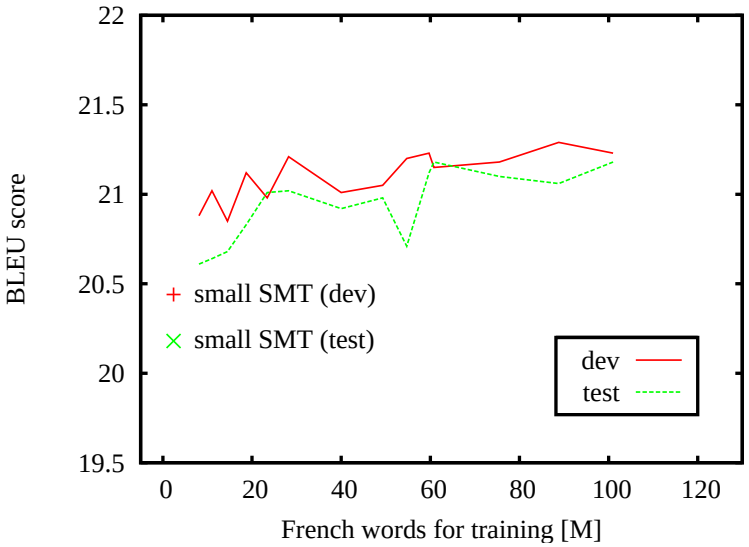
- Build SMT system with all human and automatic transl.



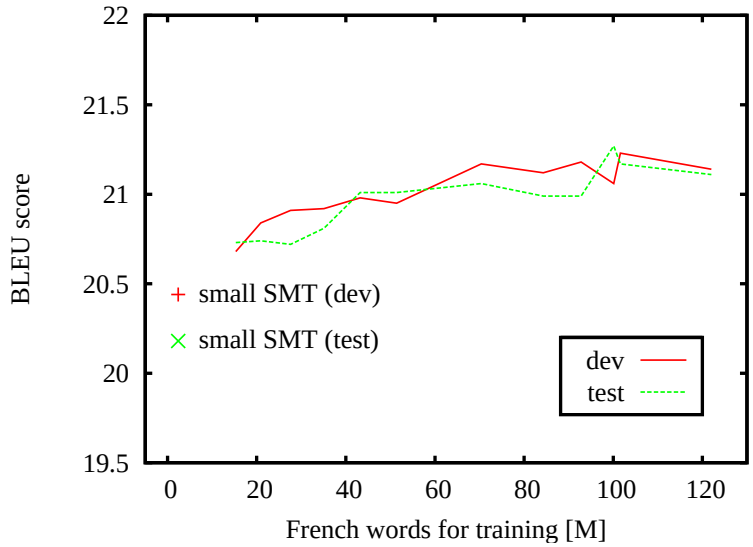
- Mainly improves performance on Test data
- Fortunate peak when using a total of 280M words of bitexts: +0.6 BLEU on Test

Using the small SMT Baseline System

Build SMT system with human-provided and translations of afp9x



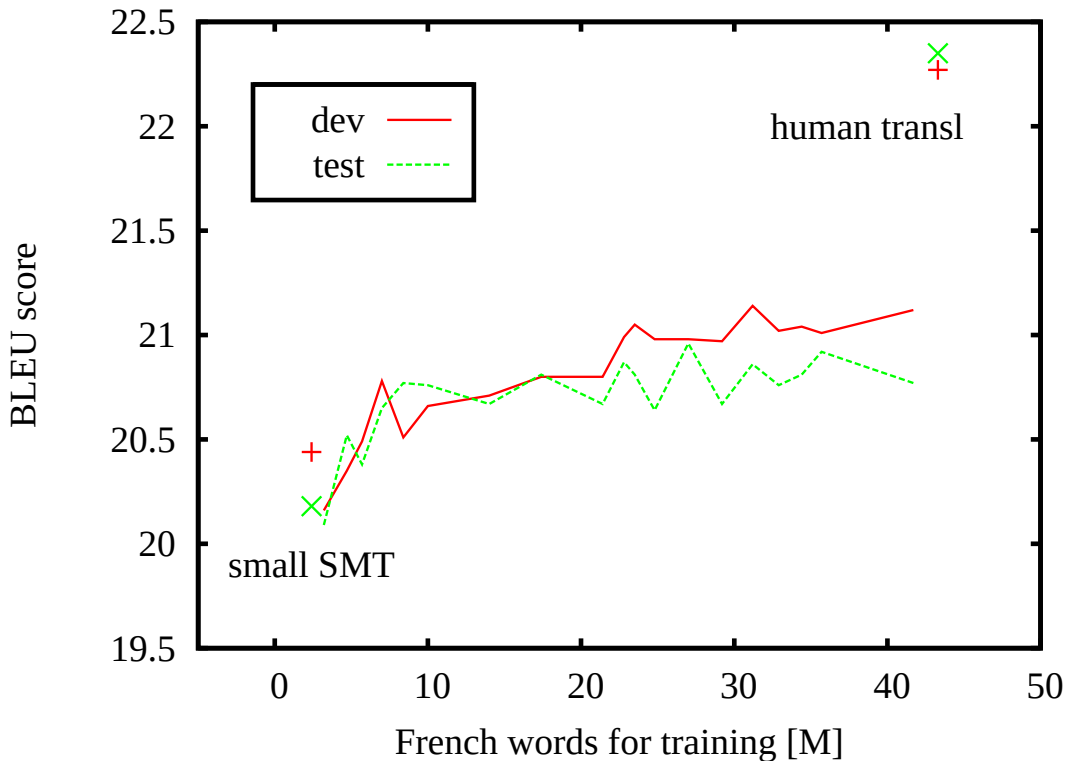
afp2x



- Best performance on Dev for a total of about 100M words
- BLEU on Test set is 21.2 (+1 point)

$\Rightarrow$ iterate the process ?

Retranslating Europarl



- Automatic translations of Europarl seem to be less useful than Gigaword data
- Comparison to the reference translations: third of the improvement with 70% of the data

Conclusions

- Translated up to 300M words from Gigaword news texts from French to English
- Automatic translation directly used as additional bitexts (after simple filtering)
- First application of large-scale lightly-supervised training to SMT
- Improvements in the BLEU score:
  - +0.6 on top of state-of-the-art system
  - +1.1 on top of small SMT system (2.4M words of bitexts)
- Seems to improve generalization behavior
- Method to obtain translation probabilities for dictionary words
- Used several thousands of hours of compute time

Perspectives

- Use a biased LM (with comparable corpora)
- Verify approach when no related texts in the target language are available
- More sophisticated techniques to filter the translations
- Iterate procedure and incrementally improve the system ?
- Compare and combine with IR techniques to extract parallel sentences
- Other language pairs

Result Summary

Bitexts Human-provided		Lightly-supervised		Total Words	BLEU score Dev	BLEU score Test	Ptable Size
Nw+dict	2.4M	-		2.4M	**20.44**	**20.18**	5M
Nw+Ep+dict	43M			43.3M	22.17	22.35	83M
Nw+Ep+Hans+dict	116M			116M	**22.69**	**22.17**	213M
Translated with the small SMT system:							
News	2.4M	afp9x	28M	2.4M	21.21	21.02	58M
			101M	2.4M	**21.23**	**21.18**	189M
		afp2x	43M	2.4M	20.98	21.01	77M
			102M	2.4M	**21.23**	**21.17**	170M
		Eparl	7M	2.4M	20.78	20.65	17M
			31M	2.4M	21.14	20.86	67M
Translated with the big SMT system:							
-		afp2x	31M	31M	22.23	22.33	55M
			112M	112M	**22.56**	**22.47**	180M
News+Eparl	42M	afp2x	77M	129M	22.65	22.44	203M
	42M		155M	197M	22.53	22.73	320M
News+Eparl+Hans	114M	afp2x	167M	281M	**22.86**	**22.80**	464M