Investigations on Translation Model Adaptation Using Monolingual Data

Patrik Lambert, Holger Schwenk
Christophe Servan and Sadaf Abdul-Rauf

LIUM (Computing Laboratory)
University of Le Mans
France

WMT 2011

Introduction

Most Statistical Machine Translation (SMT) systems rely on parallel texts

- sparse resource for most language pairs
- mostly come from particular domains
  (proceedings of the Canadian or European Parliament)
  $\Rightarrow$ problematic for general translations

Introduction

Most Statistical Machine Translation (SMT) systems rely on parallel texts

- sparse resource for most language pairs
- mostly come from particular domains
  (proceedings of the Canadian or European Parliament)
  $\Rightarrow$ problematic for general translations

Monolingual data is usually available:

- in large amounts
- in a variety of domains

Introduction

Most Statistical Machine Translation (SMT) systems rely on parallel texts

- sparse resource for most language pairs
- mostly come from particular domains
  (proceedings of the Canadian or European Parliament)
  $\Rightarrow$ problematic for general translations

Monolingual data is usually available:

- in large amounts
- in a variety of domains

$\Rightarrow$ Can we use **monolingual data** to improve somehow the translation model ?

- it's quite unlikely that we are able to introduce new translations
- but we should be able to **modify**/**adapt** the probability distributions of the existing translation model
- we may also be able to come up with new sequences of existing words and their translations

Some Background of Unsupervised Training

Large Vocabulary Speech Recognition:

- Unsupervised training is successfully used since quite some time
- Sometimes light supervison by subtitles
- Transcribe large amounts of raw audio and add the automatic transcriptions to the data (after some filtering)

Unsupervised Training in SMT

Self-Learning [Ueffing et al, IWSLT'06, ACL'07]

- Translate the **test set**, filter sentences,
- build **additional** phrase-table

Large-scale Unsupervised Training in SMT, [Schwenk, IWSLT'08]

- Use large amounts of monolingual data instead of test-set only
- Filter automatic translations using the normalised sum of the log-scores
- Use these translations instead of generic bitexts
- Build a **complete new system** using standard SMT pipeline
- French/English: improvements of about 0.6 BLEU
- Also used in Ar/Fr and Ar/En NIST and Gale systems ($\approx$ 1.0 BLEU)

Issues raised

Some open questions:

1. Choice of the translation direction: source-to-target or target-to-source
   MT is *symmetric* in contrast to ASR

2. Do we need to rebuild a system from scratch ?

3. Can we also learn new words ?

Issues raised

Some open questions:

1. Choice of the translation direction: source-to-target or target-to-source
   MT is *symmetric* in contrast to ASR

$\Rightarrow$ target-to-source is better

2. Do we need to rebuild a system from scratch ?

3. Can we also learn new words ?

Issues raised

Some open questions:

1. Choice of the translation direction: source-to-target or target-to-source
   MT is *symmetric* in contrast to ASR

⇒ target-to-source is better

2. Do we need to rebuild a system from scratch ?

⇒ No, we can re-use the alignments used during decoding
don't need to rerun giza, just construct a new phrase table

3. Can we also learn new words ?

Issues raised

Some open questions:

1. Choice of the translation direction: source-to-target or target-to-source
   MT is *symmetric* in contrast to ASR

⇒ target-to-source is better

2. Do we need to rebuild a system from scratch ?

⇒ No, we can re-use the alignments used during decoding
don't need to rerun giza, just construct a new phrase table

3. Can we also learn new words ?

⇒ use stemming to infer translations of unkown word forms in morphologically rich languages

Unsupervised Training in SMT (II)

- Ueffing et al. used later: more monolingual data, but from source language
- Chen et al, MT'08 adapt translation+language+reordering models
- Bertoldi and Federico, EACL'09
  - mention of re-use of word alignment used in decoding (very small drop in performance)
  - raise question of choice of translation direction, but seen from availability of in-domain monolingual data in source or target language
    available in source: source-to-target: adapt only TM
    available in target: target-to-source: adapt TM+LM
- Habash, ACL'08
- Bojar and Tamchyna, 2011
- Huck et al., UNSUP'11
  - use translations performed by a phrase-based system to improve a hierarchical system (cross-site adaptation)
  - improvement of about 1 point BLEU
  - It's possible to train hiero system on automatic translations **only**

Available Data

same data as those allowed for WMT 2011 shared task:

- parallel corpora:
  - Europarl + newsc : 54M words
  - Europarl + newsc + subset of 10^9 Fr/En : 285M words
- Dev=newstest2009, test=newstest2010
- LM: Gigaword + crawled news data (6.7G English / 1.5G French)

Synthetic Data

- Baseline system trained on 285M-word bitexts used for translation
- Monolingual crawled news from 2009, 2010 and 2011 were translated to adapt the systems:
  143M English words French-to-English (fe)
  248M English words English-to-French (ef)
- **after filtering**, synthetic bitext available to adapt the baseline system:
  45M English words French-to-English (fe)
  100M English words English-to-French (ef)

Synthetic Data

- Baseline system trained on 285M-word bitexts used for translation
- Monolingual crawled news from 2009, 2010 and 2011 were translated to adapt the systems:
  143M English words French-to-English (fe)
  248M English words English-to-French (ef)
- **after filtering**, synthetic bitext available to adapt the baseline system:
  45M English words French-to-English (fe)
  100M English words English-to-French (ef)
- for meaningful comparison, randomly select subset with 45M English words in English-to-French bitext

Word Alignment

Bitexts:

- baseline (manual translations: 285M word bitext)
- synthetic (automatic translations of crawled news in French)

We compare 3 word alignment configurations:

- giza: *GIZA* run on baseline+synthetic
- Could the synthetic data damage the baseline bitext alignment ?

  reused giza: *GIZA* run on baseline+synthetic, but keep orignal GIZA on baseline
- reused moses: *GIZA* on baseline + *MOSES* alignments on synthetic

Word Alignment

BLEU scores: average of 3 MERT runs (with different random seeds)
In parentheses: standard deviation

alignment	Dev	Test	
	BLEU	BLEU	TER
giza	27.34 (0.01)	29.80 (0.06)	55.34 (0.06)
reused giza	27.40 (0.05)	29.82 (0.10)	55.30 (0.02)
reused moses	27.42 (0.02)	29.77 (0.06)	55.27 (0.03)

$\Rightarrow$ no significant difference in terms of performance

Choice of Translation Direction

Le ministre de l' Intérieur tunisien est limogé .
The Minister of the Interior is Tunisian sacked .

Choice of Translation Direction

Le ministre de l' Intérieur tunisien est limogé .
The Minister of the Interior is Tunisian sacked .

⇒ malformed phrase pair:
tunisien est limogé ||| is Tunisian sacked ||| ...

- source-to-target: incorrect translations can be used in future translations
- target-to-source: incorrect translations are unlikely to match well formed input ⇒ won't be used

Translation results of the English–French systems

human translated bitexts	synthetic bitexts	Dev	Test	
		BLEU	BLEU	TER
285M	-	26.95	29.29 (0.03)	55.77 (0.19)
	fe 45M	**27.42**	**29.77 (0.06)**	**55.27 (0.03)**
	ef 45M	26.75	28.88 (0.10)	56.06 (0.05)

- adding target-to-source (fe) synthetic data:
  - 0.5 BLEU, 0.5 TER better than baseline,
- no gain when adding source-to-target (ef) synthetic data
- 54M word baseline: target-to-source also better than source-to-target
- ef and fe synthetic data are different: could this particular set of French news translated into English be more useful than the selected set of English news translated into French?
  $\Rightarrow$ to check, add same synthetic bitexts to French-to-English baseline

Translation results of the French–English systems

human translated bitexts	synthetic bitexts	Dev	Test	
		BLEU	BLEU	TER
285M	-	28.20	28.54 (0.12)	54.17 (0.15)
	fe 45M	28.02	28.40 (0.10)	54.45 (0.06)
	ef 45M	**28.24**	**28.93 (0.22)**	**53.90 (0.08)**

- adding target-to-source (ef) synthetic data:
  - 0.4 BLEU, 0.3 TER better than baseline,
- no gain when adding source-to-target (fe) synthetic data
- 54M word baseline: target-to-source also better than source-to-target

Translation results of the French–English systems

human translated bitexts	synthetic bitexts	Dev	Test	
		BLEU	BLEU	TER
285M	-	28.20	28.54 (0.12)	54.17 (0.15)
	fe 45M	28.02	28.40 (0.10)	54.45 (0.06)
	ef 45M	**28.24**	**28.93 (0.22)**	**53.90 (0.08)**

- adding target-to-source (ef) synthetic data:
  - 0.4 BLEU, 0.3 TER better than baseline,
- no gain when adding source-to-target (fe) synthetic data
- 54M word baseline: target-to-source also better than source-to-target

Conclusion

On this data set, adding synthetic data translated from target-to-source is clearly better than synthetic data translated from source-to-target

Adding more synthetic data to the French–English system

human translated bitexts	synthetic bitexts	Dev	Test	
		BLEU	BLEU	TER
285M	-	28.20	28.54 (0.12)	54.17 (0.15)
	ef 45M	28.24	28.93 (0.22)	53.90 (0.08)
	ef 65M	28.16	28.75 (0.06)	54.03 (0.14)
	ef 100M	28.28	28.96 (0.03)	53.79 (0.09)

- when adding more synthetic data, differences not greater standard variation

Analysis of French–English phrase-tables

human transl.	synthetic	entries (M)	translations	entropy
54M	-	7.16	83.83	1.84
285M	-	25.42	235.16	2.08
	fe 45M	25.54	217.21	1.81
	ef 45M	26.09	228.07	1.96
	ef 65M	26.21	226.45	1.91
	ef 100M	26.79	227.08	1.89

- adding 230M words human-translated bitext, translation options nearly multiplied by 3, entropy 1.84→2.08

Analysis of French–English phrase-tables

human transl.	synthetic	entries (M)	translations	entropy
54M	-	7.16	83.83	1.84
285M	-	25.42	235.16	2.08
	fe 45M	25.54	217.21	1.81
	ef 45M	26.09	228.07	1.96
	ef 65M	26.21	226.45	1.91
	ef 100M	26.79	227.08	1.89

- adding 230M words human-translated bitext, translation options nearly multiplied by 3, entropy 1.84→2.08
- adding 100M words of in-domain automatic translations: translation options and entropy decrease

Analysis of French–English phrase-tables

human transl.	synthetic	entries (M)	translations	entropy
54M	-	7.16	83.83	1.84
285M	-	25.42	235.16	2.08
	fe 45M	25.54	217.21	1.81
	ef 45M	26.09	228.07	1.96
	ef 65M	26.21	226.45	1.91
	ef 100M	26.79	227.08	1.89

- adding 230M words human-translated bitext, translation options nearly multiplied by 3, entropy 1.84→2.08
- adding 100M words of in-domain automatic translations: translation options and entropy decrease
- the more automatic translations added, the lower the entropy

Analysis of French–English phrase-tables

human transl.	synthetic	entries (M)	translations	entropy
54M	-	7.16	83.83	1.84
285M	-	25.42	235.16	2.08
	fe 45M	25.54	217.21	1.81
	ef 45M	26.09	228.07	1.96
	ef 65M	26.21	226.45	1.91
	ef 100M	26.79	227.08	1.89

- adding 230M words human-translated bitext, translation options nearly multiplied by 3, entropy 1.84→2.08
- adding 100M words of in-domain automatic translations: translation options and entropy decrease
- the more automatic translations added, the lower the entropy
- target-to-source automatic translations yield more translation options and a higher entropy than source-to-target

Treatment of Unknown Words

- difficulty to translate from morphologically rich languages:
  translation of words may be only known in some forms
  example: je pense (I think), tu penses (you think)
- idea [Habash, Bojar, ...]: infer possible translations from general form
  example: finies $\overset{stem}{\rightarrow}$ fini $\overset{translate}{\longrightarrow}$ finished
- Our procedure (French-to-English):
  - automatically extract a dictionary from the phrase table
  - detect unknown word
  - look for its stemmed form in the dictionary
  - propose translations based on lexical score of the phrase table

source segment	les travaux sont **finis**
stemmed	les travaux sont **fini**
proposed segment	les travaux sont <n translation="finished||ended" prob="0.008||0.0001">**finis**</n>

Treatment of Unknown Words

Problems:

- It would be good that the automatically induced translations appear in context in the phrase table
- We need to come up with meaningful translation probabilities for these entries

Use of this technique in the framework of unsupervised training

- process the unknown words while translating the monolingual data
  $\Rightarrow$ automatically induced translations of previously unknown forms will actually appear in the new adapted phrase-table

Results:

- less than 0.2% of words in test set are actually unknown
- no visible improvements in the BLEU or TER score
- we believe that this method can only improve the usability of SMT systems

Conclusion

- Unsupervised training of SMT system gets more and more popular
- Consistently improved state-of-the-art SMT systems in various domains and language pairs
- In this language pair it is clearly better to perform the automatic translations backwards
  (but one may need to build an extra system)
- There is no need to perform the word alignment step
- Interesting framework to deal with unknown word forms in morhpological rich languages

Thank you for your attention !

Currently several PhD and post-doc positions are available at LIUM
(please contact Holger Schwenk)