Applications of Data Selection via Cross-Entropy Difference for Real-World Statistical Machine Translation

Amittai Axelrod
QingJun Li
William Lewis

Microsoft Research

Microsoft Translator

Data Selection in lieu of Domain Adaptation

- Domain adaptation:
  - Build system over (all?) available training data
  - Adjust for new task
- Cons:
  - Large systems are expensive!
  - Out-of-domain systems aren't great!
- Goal:
  - Task-specific system that is better than the (unadjusted) full system.

Data Selection in lieu of Domain Adaptation

- Data Selection:
  different way of reaching similar goal.
- If the target task is known:
  - Identify the most relevant parts of training data.
  - Build a system on only this subset.
- Goal:
  - Task-specific system >> (unadjusted) full system.
  - Task-specific system >> **adjusted** full system, too!

Some Methods for Domain Adaptation

- Multiple Translation Models
  - Drexler et al
  - Peitz et al
- Phrase-table interpolation/fill-up
  - Ruiz et al
- Multiple translation models
  - Hasler et al
- Instance reweighting
  - Mansour & Ney
- Factored RNNLMs
  - Yamamoto et al

Cross-Entropy Difference

- Leverage the fact that the data pool does not match the target task [Moore, Lewis 2010].
- Score and rank by cross-entropy difference:

$$\underset{s\ \in POOL}{\operatorname{argmin}} \quad H_{LM_{IN}}(s) - H_{LM_{POOL}}(s)$$

- Biases towards sentences that are:
  – Like the target task
  – Unlike the pool average.

What’s Wrong?

- Using BTEC data as in-domain for Chinese-English, apply data selection methods:

System	BTEC dev	BTEC test	Translation Model	Language Model
BTEC	21.68	17.02	BTEC	BTEC
data-MSR	20.88	16.37	General (bilingual)	General (bilingual)
select M-L (10%)	22.21	17.23	Selected Data 1.3m	Selected Data 1.3m

- Data selection methods can be a cheap substitute for domain adaptation [EMNLP ‘11]

What’s Wrong?

- We also looked at another test set from an online hotel review (OHR) site:

System	BTEC dev	BTEC test	OHR	Translation Model	Language Model
BTEC	21.68	17.02	4.89	BTEC	BTEC
data-MSR	20.88	16.37	15.05	General (bilingual)	General (bilingual)
select ML (10%)	22.21	17.23	10.09	Selected Data 1.3m	Selected Data 1.3m

- Real-world goal:
  The adapted system must do well on the target set…
  and still do OK on everything else

Is the Task to Blame?

- BTEC:
  Great resource for specific scenario.
- However users mis-use everything!
- Broaden the travel domain to include guidebooks, travel reviews, hotel information, brochures, etc.
- Unified but unconstrained travel task

Data Selection Survey Work

- Questions to answer:
  - – Best strategy to build travel domain systems?
    - Mono vs. bilingual data selection?
    - Build standalone travel systems?
    - Use travel domain dev data to tune general system?
  - – Increase typological/data diversity:
    Spanish, Hebrew, Czech ←→ English
    Does that affect selection effectiveness?
  - – Is there a unified strategy across language pairs?

Data

- English – **Hebrew**
  - 74k parallel in-domain
  - 3m parallel non-specific
- English – **Czech**
  - 129k parallel in-domain
  - 11m parallel non-specific
- English – **Spanish**
  - 4k parallel in-domain
  - 25m parallel non-specific
- English
  - 600k monolingual in-domain

Systems

We built the following for each language pair:

System	Dev Set	TM 0	TM 1	LM 0	LM 1
Baseline	General	General	--	All-Mono	--
Adapted Baseline	Travel	General	--	All-Mono	--
Top 10%	Travel	Top 10%	--	Top 10%	--
Top TM, All-Mono LM	Travel	Top 10%	--	All-Mono	--
Top + All-Mono LM	Travel	Top 10%		Top 10%	All-Mono
Augmented	Travel	Top 10%	General	Top 10%	All-Mono

Hebrew-English

System EN -> HE	Dev Set	TM 0	TM 1	LM 0	LM 1	Guidebook	WMT 2009
Baseline	User logs	General	--	All-Mono	--	12.04	14.88
Adapted Baseline	Travel	General	--	All-Mono	--	12.45	14.53
Augmented Bi M-L	Travel	Top 10%	General	Top 10%	All-Mono	13.49	13.84

System HE -> EN	Dev Set	TM 0	TM 1	LM 0	LM 1	Guidebook	WMT 2009
Baseline	User logs	General	--	All-Mono	--	18.18	25.03
Adapted Baseline	Travel	General	--	All-Mono	--	18.58	25.18
Augmented Mono M-L	Travel	Top 10%	General	Top 10%	All-Mono	19.12	24.92

Czech-English

System EN -> CZ	Dev Set	TM 0	TM 1	LM 0	LM 1	Guidebook	WMT 2010
Baseline	WMT	General	--	All-Mono	--	27.33	15.59
Adapted Baseline	Travel	General	--	All-Mono	--	27.73	15.03
Augmented Bi M-L	Travel	Top 10%	General	Top 10%	All-Mono	27.80	14.88

System CZ -> EN	Dev Set	TM 0	TM 1	LM 0	LM 1	Guidebook	WMT 2010
Baseline	WMT	General	--	All-Mono	--	32.52	23.88
Adapted Baseline	Travel	General	--	All-Mono	--	34.06	21.83
Augmented Bi M-L	Travel	Top 10%	General	Top 10%	All-Mono	35.48	22.15

Spanish-English

System EN-> ES	Dev Set	TM 0	TM 1	LM 0	LM 1	Travel Reviews	Hotel Reviews	WMT 2010
Baseline	WMT	General	--	All-Mono	--	32.28	29.09	32.21
Adapted Baseline	Travel	General	--	All-Mono	--	33.27	28.19	31.00
Augmented M-L	Travel	Top 10%	General	Top 10%	General	33.55	28.80	30.81

System ES-> EN	Dev Set	TM 0	TM 1	LM 0	LM 1	Travel Reviews	Hotel Reviews	WMT 2010
Baseline	WMT	General	--	All-Mono	--	38.71	32.03	32.11
Adapted Baseline	Travel	General	--	All-Mono	--	39.43	32.79	31.38
Augmented M-L	Travel	Top 10%	General	Top 10%	General	40.00	33.28	31.05

Summary

- Data selection helps even compared against production-sized SMT systems!
- In-domain performance gain > general-domain loss.
- Can improve in-domain performance without:
  - True in-domain translation system
  - Bilingual in-domain data
- Low difference between monolingual and bilingual data selection *when one language is morphologically simple*.

Questions?

The Language/Data Landscape

Language	Available Resources			Available Dev			Available Test			
	600K ENU mono	BTEC parallel	Guidebook Parallel Train	BTEC dev	Online Travel Review Dev	Guidebook Dev	BTEC Test	OTR Test	Guidebook Test	Online Hotel Review Test
CHS	X	30K		X			X			972
ESN	X				2930			776		972
CSY	X		141922			1984			4844	
HEB	X		81905			1979			4764	