A Simple and Effective Weighted Phrase Extraction for Machine Translation Adaptation

Saab Mansour, Hermann Ney

<surname>@cs.rwth-aachen.de

IWSLT 2012, Hong Kong
December 7, 2012

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Agenda

- Motivation
- Related work
- Weighted phrase extraction framework
- Weights estimation
- Experimental setup and results
- Translation examples
- Mixture modeling and results
- Conclusions and outlook

Motivation

- New domains of translation are emerging due to the success of SMT
  - e.g. lectures, patents, forums, sms
- Small in-domain (IN) training data
- Large amounts of data were already collected for other domains (OD)
  - e.g. news, government/parliamentary documents
- Domain-adaptation to utilize existing resources beneficially for new domain

Related Work

- **Sentence filtering: [Axelrod & He+ 11] - University of Washington**
  - **translation model (TM) adaptation using LM cross-entropy difference of both source and target sides**
  - **smaller TM, but discards training sentences**
- **Sentence weighting: [Matsoukas & Rosti+ 09] - BBN**
  - **discriminative training of weights optimized on the development set**
  - **details: need to define features (meta-data), mapping from features to weights (perceptron), mapping parameters are optimized using LBFGS minimizing expected TER on dev (iterated)**
- **Phrase weighting: [Foster & Goutte+ 10] - NRC**
  - **phrase level weighting, using MLE criterion on development**
  - **using the weights to directly model phrase probabilities underperforms weighting**

Weighted Phrase Extraction Framework

- **Phrase model estimated using relative frequency:**

$$p(\tilde{f}|\tilde{e}) = \frac{\sum_r c_r(\tilde{f},\tilde{e})}{\sum_{\tilde{f}'}\sum_r c_r(\tilde{f}',\tilde{e})}$$

  - $\tilde{f},\tilde{e}$: **contiguous phrases,** $c_r(\tilde{f},\tilde{e})$: **count of** $(\tilde{f},\tilde{e})$ **being translation of each other in sentence pair** (s_r, t_r)

- **Introduce weights:**

$$p(\tilde{f}|\tilde{e}) = \frac{\sum_r w_r \cdot c_r(\tilde{f},\tilde{e})}{\sum_{\tilde{f}'}\sum_r w_r \cdot c_r(\tilde{f}',\tilde{e})}$$

Weights Estimation

- Ad-hoc (similar to GIZA++), e.g., higher weight to in-domain (10*IN+1*OD)
- [Axelrod et al., 2011] use LM perplexity for bilingual corpora filtering
- Utilize same scoring method, but for weighting:

$$d_r = [H_{LM_{IN,src}}(s_r) - H_{LM_{OD,src}}(s_r)] + [H_{LM_{IN,trg}}(t_r) - H_{LM_{OD,trg}}(t_r)]$$

$$w_r = e^{-d_r}$$

  - $H_{LM}(s)$: cross-entropy of sentence s according to LM
  - $H_{LM_{IN}}$ smaller is closer to IN
  - $H_{LM_{OD}}$ bigger is further from OD
  - $\Rightarrow [H_{LM_{IN}} - H_{LM_{OD}}]$ smaller is closer to IN and further from OD
  - $\Rightarrow d_r$ smaller is better, w_r bigger is better
- Compare above weighting to:
  - source only LM ppl-src: $d_r = [H_{LM_{IN,src}}(s_r) - H_{LM_{OD,src}}(s_r)]$
  - target only LM ppl-trg: $d_r = [H_{LM_{IN,trg}}(t_r) - H_{LM_{OD,trg}}(t_r)]$

Experiment Setup: BOLT P1

Data style	Sentences	Tokens
United Nations	3557K	122M
Newswire	1918K	57M
Web	13K	280K
Newsgroup	25K	720K
Broadcast	91K	2M
Lexicons	213K	530K
Iraqi, Levantine	617K	4M
General (sum of above)	6434K	187M
Egyptian	240K	3M

- **LM is trained over 8 billion words (4B words from the LDC gigaword corpus and 4B words collected from web resources)**

BOLT P1: OOV rates

Set	Sentences	Tokens	OOV/IN	OOV/ALL
Egyptian (IN)	240K	3M		
General (OD)	6.4M	187M		
dev	1219	18K	387 (2.2%)	160 (0.9%)
test	1510	27K	559 (2.1%)	201 (0.7%)

- General set considerably reduces the number of OOV words
- Increasing size of the training data by a factor of more than 50
- Simple concatenation of the corpora might mask the IN phrase probabilities
- Filtering the general corpus: discards phrase translations completely
- Weighting: sentences more related to the domain will have higher weights

Results

Adaptation	Translation model	dev BLEU	dev TER	test BLEU	test TER
None	IN	24.6	61.2	22.2	62.6
	IN+OD	25.3	60.6	22.5	61.9
Filtering	IN+OD-0.25Mbest	25.1	61.0	22.3	62.5
	IN+OD-0.5Mbest	25.1	60.6	22.5	61.8
	IN+OD-1Mbest	25.4	60.5	22.9	61.6
	IN+OD-2Mbest	25.4	60.5	22.6	61.8
Weighting	10IN+1OD	25.6	60.2	22.8	61.5
	ppl-src(IN+OD)	25.6	60.6	23.3	61.0
	ppl-trg(IN+OD)	25.6	60.6	22.8	61.8
	ppl(IN+OD)	25.6	60.1	23.3	60.9
Filtering+Weighting	ppl(IN+OD-1Mbest)	25.6	60.0	23.0	61.4

- OD data is useful, filtering improves the results
- Ad-hoc weighting does not improve over filtering
- Weighting improves results, best is by using both source and target LMs
- Filtering+weighting is useful to reduce TM size and get best results

Phrase Table Examples

Arabic $\tilde{f}$	IN+OD $\tilde{e}$	IN+OD $-\log(p)$	ppl(IN+OD) $\tilde{e}$	ppl(IN+OD) $-\log(p)$
الميدان [AlmydAn]	field	3.1	the square	1.7
	the field	4.2	field	3.7
	the square	5.1	the square ,	4.2
	the ground	6.2	the field	4.4
خلي بال +ك [xly bAl +k]	watch out	4.7	watch out	3.9
	let your mind	5.3	be careful	4.6
	watch out you	5.3	take care	4.9
	watch out your	5.7	watch out you	5.2

- Comparing unweighted (IN+OD) and weighted (ppl(IN+OD)) tables
- Entries sorted by phrase probability ($-log(p(\tilde{f}|\tilde{e}))$)

Translation Examples

Source	و مرحت +ش المظاهرات في الميدان
Buckwalter	**w+ mrht +$ AlmZAhrAt fy AlmydAn**
Reference	**but i didn’t go to the demonstrations in the square ,**
IN+OD	**and i didn’t demonstrations in the field .**
ppl(IN+OD)	**and i didn’t demonstrations in the square .**

Source	و خلي بال +ك السلطة روجت قبل كده ...
Buckwalter	**w+ xly bAl +k AlslTp rwjt qbl kdh ...**
Reference	**but remember , that the authority said earlier ...**
IN+OD	**and let your mind power promoted before**
ppl(IN+OD)	**and take care of the authority promoted before**

Mixture Modeling

- **Mixture modeling can be used for adaptation purposes**
- **Compare weighting to mixture modeling, and combine the methods**
- **Linear interpolation:** $p(\tilde{f}|\tilde{e}) = \lambda p_{IN}(\tilde{f}|\tilde{e}) + (1-\lambda) p_{OD}(\tilde{f}|\tilde{e})$**, manual optimization of** λ
- **Loglinear interpolation: fits directly into the SMT loglinear framework, weights optimized using MERT**
- **ifelse method: if phrase pair exists in IN, use IN probability, otherwise use OD probability [Haddow & Koehn, 12] (fill-up)**

Mixture Modeling - Results

Translation model	dev BLEU	dev TER	test BLEU	test TER
Unfiltered				
IN	24.6	61.2	22.2	62.6
IN+OD	25.3	60.6	22.5	61.9
Weighted phrase extr.				
ppl(IN+OD)	25.6	60.1	23.3‡	60.9‡
Mixture modeling				
IN-loglin-IN+OD	24.7	61.3	22.0	62.8
IN-loglin-ppl(IN+OD)	24.9	61.1	22.1	62.3
IN-linear-IN+OD	25.7	60.4	22.9	61.4
IN-linear-ppl(IN+OD)	26.0	59.9	23.3‡	60.6‡
IN-ifelse-IN+OD	25.6	60.2	23.0	61.1
IN-ifelse-ppl(IN+OD)	25.7	60.2	23.1	61.0

- loglinear interpolation hinders performance
- linear interpolation with weighted ppl(IN+OD) performs best
- ifelse combination is competitive while simple

Conclusions and Outlook

- **Conclusions:**
  - **introduced a general framework for weighted phrase extraction**
  - **applied for adaptation, significant improvements**
  - **compared weighting to recent work: filtering and mixture modeling**
  - **better results for weighting, weighting+mixture yields further improvements**
- **Outlook:**
  - **compare different weighting methods**
  - **compare different granularity**
  - **apply to other models: lexical smoothing, lexicalized reordering...**

Thank you for your attention

Saab Mansour

`mansour@cs.rwth-aachen.de`

`http://www-i6.informatik.rwth-aachen.de/`