Bayesian Reordering Model with Feature Selection

Abdullah Alrajeh and **Mahesan Niranjan**

The Ninth Workshop on Statistical Machine Translation
Baltimore, Maryland USA
June 26-27, 2014

Translation System Overview

Given a foreign sentence **f**, the best translation **e** is (Brown et al., 1993):

$$\mathbf{e}_{\text{best}} = \arg\max_{\mathbf{e}} \; p(\mathbf{e}|\mathbf{f})$$

$$= \arg\max_{\mathbf{e}} \; \frac{p(\mathbf{f}|\mathbf{e})\,p(\mathbf{e})}{p(\mathbf{f})}$$

$$= \arg\max_{\mathbf{e}} \; \text{Translation Model} \times \text{Language Model}$$

Translation System Overview

Given a foreign sentence **f**, the best translation **e** is (Brown et al., 1993):

$$\mathbf{e}_{\text{best}} = \arg\max_{\mathbf{e}} \; p(\mathbf{e}|\mathbf{f})$$
$$= \arg\max_{\mathbf{e}} \; \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{p(\mathbf{f})}$$
$$= \arg\max_{\mathbf{e}} \; \text{Translation Model} \times \text{Language Model}$$

$$\mathbf{e}_{\text{best}} = \arg\max_{\mathbf{e}} \left\{ p_t(\mathbf{f}|\mathbf{e})^{\lambda_t} p_{lm}(\mathbf{e})^{\lambda_{lm}} p_{lex}(\mathbf{f}|\mathbf{e})^{\lambda_{lex}} p_{reo}(\mathbf{f},\mathbf{e})^{\lambda_{reo}} w^{|\mathbf{e}|\lambda_w} \right\}$$
$$= \arg\max_{\mathbf{e}} \sum_i \lambda_i \log p_i(\mathbf{f},\mathbf{e})$$

Translation System Overview

Given a foreign sentence **f**, the best translation **e** is (Brown et al., 1993):

$$\mathbf{e}_{\text{best}} = \arg\max_{\mathbf{e}} \; p(\mathbf{e}|\mathbf{f})$$
$$= \arg\max_{\mathbf{e}} \; \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{p(\mathbf{f})}$$
$$= \arg\max_{\mathbf{e}} \; \text{Translation Model} \times \text{Language Model}$$

$$\mathbf{e}_{\text{best}} = \arg\max_{\mathbf{e}} \left\{ p_t(\mathbf{f}|\mathbf{e})^{\lambda_t} p_{lm}(\mathbf{e})^{\lambda_{lm}} p_{lex}(\mathbf{f}|\mathbf{e})^{\lambda_{lex}} p_{reo}(\mathbf{f},\mathbf{e})^{\lambda_{reo}} w^{|\mathbf{e}|\lambda_w} \right\}$$
$$= \arg\max_{\mathbf{e}} \sum_i \lambda_i \log p_i(\mathbf{f},\mathbf{e})$$

In general, reordering model is defined as:

$$p_{reo}(\mathbf{f},\mathbf{e}) = \prod_n p(o_n|\bar{f}_n, \bar{e}_n) = \prod_n \frac{h(\bar{f}_n, \bar{e}_n, o_n)}{\sum_k h(\bar{f}_n, \bar{e}_n, o_k)}$$

Reordering Models

Foreign sentence $\mathbf{f}$: $\bar{f}_1$ $\bar{f}_2$ $\bar{f}_3$.

English sentence $\mathbf{e}$: $\bar{e}_1$ $\bar{e}_3$ $\bar{e}_2$.

$$p_{reo}(\mathbf{f}, \mathbf{e}) = p(o_1 = \text{mono}|\bar{f}_1, \bar{e}_1) \times p(o_2 = \text{swap}|\bar{f}_2, \bar{e}_2) \times p(o_3 = \text{other}|\bar{f}_3, \bar{e}_3)$$

Reordering Models

$$\text{Foreign sentence } \mathbf{f}: \quad \bar{f}_1 \quad \bar{f}_2 \quad \bar{f}_3 \; .$$

$$\text{English sentence } \mathbf{e}: \quad \bar{e}_1 \quad \bar{e}_3 \quad \bar{e}_2 \; .$$

$$p_{reo}(\mathbf{f}, \mathbf{e}) = p(o_1 = \text{mono} | \bar{f}_1, \bar{e}_1) \times p(o_2 = \text{swap} | \bar{f}_2, \bar{e}_2) \times p(o_3 = \text{other} | \bar{f}_3, \bar{e}_3)$$

- Lexicalized Reordering Model (Tillmann, 2004; Kumar and Byrne, 2005; Koehn et al., 2005; Galley and Manning, 2008)

$$p(o_k | \bar{f}_n, \bar{e}_n) = \frac{\text{count}(\bar{f}_n, \bar{e}_n, o_k)}{\sum_{k'} \text{count}(\bar{f}_n, \bar{e}_n, o_{k'})}$$

Reordering Models

Foreign sentence $\mathbf{f}$: $\bar{f}_1$ $\bar{f}_2$ $\bar{f}_3$.

English sentence $\mathbf{e}$: $\bar{e}_1$ $\bar{e}_3$ $\bar{e}_2$.

$$p_{reo}(\mathbf{f},\mathbf{e}) = p(o_1 = \text{mono}|\bar{f}_1,\bar{e}_1) \times p(o_2 = \text{swap}|\bar{f}_2,\bar{e}_2) \times p(o_3 = \text{other}|\bar{f}_3,\bar{e}_3)$$

- Lexicalized Reordering Model (Tillmann, 2004; Kumar and Byrne, 2005; Koehn et al., 2005; Galley and Manning, 2008)

$$p(o_k|\bar{f}_n,\bar{e}_n) = \frac{\text{count}(\bar{f}_n,\bar{e}_n,o_k)}{\sum_{k'} \text{count}(\bar{f}_n,\bar{e}_n,o_{k'})}$$

- Discriminative Reordering Model (Zens and Ney, 2006; Xiong et al., 2006; Nguyen et al., 2009; Xiang et al., 2011; Ni et al., 2011)

$$p(o_k|\bar{f}_n,\bar{e}_n) = \frac{\exp(\mathbf{w}_k^T\phi(\bar{f}_n,\bar{e}_n))}{\sum_{k'}\exp(\mathbf{w}_{k'}^T\phi(\bar{f},\bar{e}))} \equiv \frac{\exp(\mathbf{w}^T\phi(\bar{f}_n,\bar{e}_n,o_k))}{\sum_{k'}\exp(\mathbf{w}^T\phi(\bar{f},\bar{e},o_{k'}))}$$

Feature Extraction

Foreign sentence **f** : $\underbrace{f_1 \quad f_2}_{1} \quad \underbrace{f_3 \quad f_4 \quad f_5}_{2} \quad \underbrace{f_6}_{3}$.

English sentence **e** : $\underbrace{e_1}_{1} \quad \underbrace{e_2 \quad e_3}_{3} \quad \underbrace{e_4 \quad e_5}_{2}$.

Feature Extraction

Foreign sentence **f** : $[f_1 \; f_2]_1 \; [f_3 \; f_4 \; f_5]_2 \; [f_6]_3$.

English sentence **e** : $[e_1]_1 \; [e_2 \; e_3]_3 \; [e_4 \; e_5]_2$.

Extracted phrase pairs :

$\bar{f}_n$	|||	$\bar{e}_n$	|||	o_n	|||	word alignment	|||	context feature
$f_1 \; f_2$	|||	e_1	|||	mono	|||	0-0 1-0	|||	$+f_3$
$f_3 \; f_4 \; f_5$	|||	$e_4 \; e_5$	|||	swap	|||	0-1 2-0	|||	$-f_2 \; +f_6$
f_6	|||	$e_2 \; e_3$	|||	other	|||	0-0 0-1	|||	$-f_5$

Feature Extraction

Foreign sentence **f** : $\boxed{f_1 \ f_2}_1 \ \ \boxed{f_3 \ f_4 \ f_5}_2 \ \ \boxed{f_6}_3$.

English sentence **e** : $\boxed{e_1}_1 \ \ \boxed{e_2 \ e_3}_3 \ \ \boxed{e_4 \ e_5}_2$.

Extracted phrase pairs :

$\bar{f}_n$	|||	$\bar{e}_n$	|||	o_n	|||	word alignment	|||	context feature
$f_1 \ f_2$	|||	e_1	|||	mono	|||	0-0 1-0	|||	$+f_3$
$f_3 \ f_4 \ f_5$	|||	$e_4 \ e_5$	|||	swap	|||	0-1 2-0	|||	$-f_2 \ + f_6$
f_6	|||	$e_2 \ e_3$	|||	other	|||	0-0 0-1	|||	$-f_5$

All linguistic features:

$(f_1\&e_1)^1 \ (f_2\&e_1)^2 \ (+f_3)^3 \ (f_3\&e_5)^4 \ (f_5\&e_4)^5 \ (-f_2)^6 \ (+f_6)^7 \ (f_6\&e_2)^8 \ (f_6\&e_3)^9 \ (-f_5)^{10}$

Bag-of-words representation (0=not exist):

$\phi(\bar{f}_n, \bar{e}_n)$	1	2	3	4	5	6	7	8	9	10
$\phi(\bar{f}_1, \bar{e}_1)$ =	1	1	1	0	0	0	0	0	0	0
$\phi(\bar{f}_2, \bar{e}_2)$ =	0	0	0	1	1	1	1	0	0	0
$\phi(\bar{f}_3, \bar{e}_3)$ =	0	0	0	0	0	0	0	1	1	1

The Proposed Reordeing Model

Naive Bayes

$$p(o_k|\bar{f}_n, \bar{e}_n) = \frac{p(\bar{f}_n, \bar{e}_n|o_k)p(o_k)}{\sum_{k'} p(\bar{f}_n, \bar{e}_n|o)p(o_{k'})}.$$

Multinomial distribution:

$$p(\bar{f}_n, \bar{e}_n|\mathbf{q}_k) = C \prod_m^M {q_{km}}^{\phi_m(\bar{f}_n, \bar{e}_n)}$$

Naive Bayes

$$p(o_k|\bar{f}_n, \bar{e}_n) = \frac{p(\bar{f}_n, \bar{e}_n|o_k)p(o_k)}{\sum_{k'} p(\bar{f}_n, \bar{e}_n|o)p(o_{k'})}.$$

Multinomial distribution:

$$p(\bar{f}_n, \bar{e}_n|\mathbf{q}_k) = C \prod_m^M q_{km}{}^{\phi_m(\bar{f}_n, \bar{e}_n)}$$

Maximum-likelihood estimation:

$$q^*_{km} = \arg\max_{\mathbf{q}_k} \prod_n^{N_k} p(\bar{f}_n, \bar{e}_n|\mathbf{q}_k) = \frac{\sum_n^{N_k} \phi_m(\bar{f}_n, \bar{e}_n)}{\sum_{m'}^M \sum_n^{N_k} \phi_{m'}(\bar{f}_n, \bar{e}_n)}.$$

The Proposed Reordeing Model

Naive Bayes

$$p(o_k|\bar{f}_n,\bar{e}_n) = \frac{p(\bar{f}_n,\bar{e}_n|o_k)p(o_k)}{\sum_{k'} p(\bar{f}_n,\bar{e}_n|o)p(o_{k'})}.$$

Multinomial distribution:

$$p(\bar{f}_n,\bar{e}_n|\mathbf{q}_k) = C\prod_{m}^{M} {q_{km}}^{\phi_m(\bar{f}_n,\bar{e}_n)}$$

Maximum-likelihood estimation:

$$q^*_{km} = \arg\max_{\mathbf{q}_k} \prod_{n}^{N_k} p(\bar{f}_n,\bar{e}_n|\mathbf{q}_k) = \frac{\sum_n^{N_k}\phi_m(\bar{f}_n,\bar{e}_n)}{\sum_{m'}^{M}\sum_n^{N_k}\phi_{m'}(\bar{f}_n,\bar{e}_n)}.$$

Maximum a posteriori (MAP) estimation:

$$q^*_{km} = \arg\max_{\mathbf{q}_k} \prod_{n}^{N_k} p(\bar{f}_n,\bar{e}_n|\mathbf{q}_k)p(\mathbf{q}_k|\alpha) = \frac{\alpha - 1 + \sum_n^{N_k}\phi_m(\bar{f}_n,\bar{e}_n)}{M(\alpha-1)+\sum_{m'}^{M}\sum_n^{N_k}\phi_{m'}(\bar{f}_n,\bar{e}_n)}$$

The Proposed Reordeing Model

Bayesian Naive Bayes (Barber, 2012)

$$p(o_k|\bar{f}_n,\bar{e}_n) = \frac{p(\bar{f}_n,\bar{e}_n|o_k)p(o_k)}{\sum_{k'} p(\bar{f}_n,\bar{e}_n|o)p(o_{k'})}.$$

Full Bayesian Inference

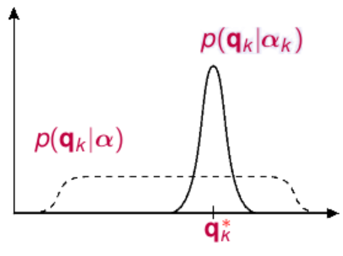
Multinomial-Dirichlet

$$p(\bar{f}_n,\bar{e}_n|o_k) = \int p(\bar{f}_n,\bar{e}_n|\mathbf{q}_k)p(\mathbf{q}_k|\alpha_k)\,\mathrm{d}\mathbf{q}_k$$

$$= C\,\frac{\Gamma\left(\sum_m \alpha_{km}\right)}{\prod_m \Gamma(\alpha_{km})}\,\frac{\prod_m \Gamma(\alpha_{km}+\phi_m(\bar{f}_n,\bar{e}_n))}{\Gamma\left(\sum_m \alpha_{km}+\phi_m(\bar{f}_n,\bar{e}_n)\right)}$$

$$p(\mathbf{q}_k|\alpha_k) = \frac{p(\mathbf{q}_k|\alpha)\prod_n^{N_k} p(\bar{f}_n,\bar{e}_n|\mathbf{q}_k)}{\int p(\mathbf{q}_k|\alpha)\prod_n^{N_k} p(\bar{f}_n,\bar{e}_n|\mathbf{q}_k)\mathrm{d}\mathbf{q}_k}$$

Prior and Posterior



$$\boldsymbol{\alpha}_k = \boldsymbol{\alpha} + \sum_n^{N_k} \phi(\bar{f}_n,\bar{e}_n)$$

Classification Results

3-class problem: mono , swap , other

Table: Arabic-English MultiUN corpus (Eisele and Chen, 2010)

Statistics	Arabic	English
Sentence Pairs	9.7 M	
Running Words	255.5 M	285.7 M
Word/Line	22	25
Vocabulary Size	677 K	410 K

Classification Results

3-class problem: mono , swap , other

Table: Arabic-English MultiUN corpus (Eisele and Chen, 2010)

Statistics	Arabic	English
Sentence Pairs	9.7 M	
Running Words	255.5 M	285.7 M
Word/Line	22	25
Vocabulary Size	677 K	410 K

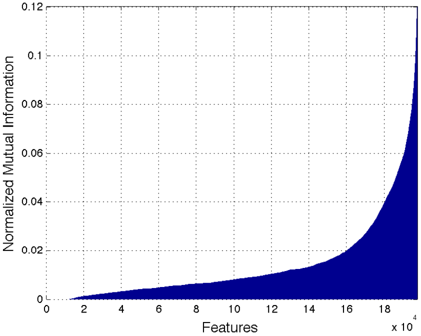
Table: Error rate based on 3-fold cross-validation

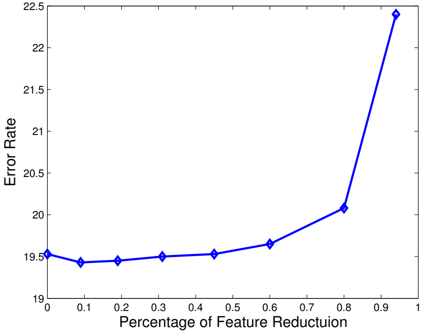
Classifier	Error Rate
Lexicalized model	25.2%
Bayes-MAP estimate	**19.53%**
Bayes-Bayesian inference	20.13%

Feature Selection

Normalized mutual information (Estevez et al., 2009):

$$I_{norm}(X;Y) = \frac{I(X;Y)}{\min(H(X),H(Y))}.$$





Translation Results

Table: NIST test sets (4 references for each Arabic sentence)

Evaluation Set		Arabic	English
NIST MT06	sentences	1797	7188
	words	49 K	223 K
NIST MT08	sentences	813	3252
	words	25 K	117 K

Translation Results

Table: NIST test sets (4 references for each Arabic sentence)

Evaluation Set		Arabic	English
NIST MT06	sentences	1797	7188
	words	49 K	223 K
NIST MT08	sentences	813	3252
	words	25 K	117 K

Table: BLEU Score (Papineni et al., 2002)

Translation System	ReoM Size	Speed	MT06	MT08
Baseline	-	-	28.92	32.13
BL + Lexicalized ReoM	604 MB	2.2 sec/s	30.86	34.22
BL + Bayes-MAP ReoM	18 MB	2.6 sec/s	**31.21**	**34.72**
BL + Bayes-Baysien ReoM	18 MB	2.6 sec/s	**31.20**	**34.69**

Thank you for your attention.