Evaluation in Machine Translation

Emine Sakir, Stefan Petrik

Overview

Problems of the N-Gram Approach

Word Error Rate (WER) Based Measures

- WER, mWER (Word Error Rate, multi-reference Word Error Rate)
- PER, mPER (Position-independent word Error Rate, multi-reference)
- GSA (Generation String Accuracy)
- RED (gRader based on Edit Distances)

Minimum Error Rate Training

Word Error Rate (WER) Based Measures

Problems of the n-gram approach

- position-dependent score
- intolerance towards small errors as in conversational speech

Problems of the N-Gram Approach

Position-dependent score

```
   I brought a small white flower to my girl.
1) I took    a small white flower to my girl.
2) I once brought a small white flower to my girl.
3) I _______ a  small white flower to my girl.
```

```
   I brought a small  white flower to my girl.
1) I brought a little white flower to my girl.
2) I brought a very small white flower to my girl.
3) I brought a _____  white flower to my girl.
```

Position i	1	2	3	4	5	6	7	8	9
(Case 1)	1/6	2/6	3/6	4/6	4/6	4/6	3/6	2/6	1/6
(Case 2)	1/7	2/7	3/7	4/7	4/7	4/7	4/7	3/7	2/7
(Case 3)	0/5	1/5	2/5	3/5	4/5	3/5	2/5	1/5	0/5

Problems of the N-Gram Approach

Intolerance for small deviations

- word swap
- semantically similar words
- differentiation between content & function words

Example

I brought a small white flower to my girl.

1) I brought a white small flower to my girl.
2) I brought a little snow-white flower to my girl.
3) I brought small white flower my girl.
4) I brought a flower to my girl.

Word Error Rate (WER) Based Measures

WER (Word Error Rate)

- sum of substitutions (S), insertions (I), and deletions (D) between machine-translated text and reference translation in relation to number of words in reference translation
- multiple references: select minimum WER

$$WER = \frac{(S+I+D)}{R} \qquad mWER = min_i \frac{(S_i+I_i+D_i)}{R_i}$$

PER (Position-independent word Error Rate)

- sentence = bag of words (no word positions)
- *PER = number of differences between machine-translated text and reference translation*

GSA (Generation String Accuracy)

- consider moves *M* (=*ins*+*del* of same element) as one edit operation

$$GSA = 1 - \frac{(M+S+I'+D')}{N}$$

Word Error Rate (WER) Based Measures

Examples

- Ref = w_1 w_2 w_3
  MT = w_1 w_3 w_2 w_4

  WER = 2/3 (1 INS, 1 SUB)
  PER = 1/3 (1 SUB)
  GSA = 1/3 (1 MOV, 1 INS)

- Ref = w_1 w_2 w_3 w_4
  MT = w_2 w_3 w_4 w_1

  WER = 2/4 (1 INS, 1 DEL)
  PER = 0
  GSA = 3/4 (1 MOV)

Word Error Rate (WER) Based Measures

RED (gRader based on Edit Distances)

Idea

- learn human judgement from small set of sample human gradings
- use multiple edit distances as features
- reduce complexity of grading task to grading scale A,B,C,D

Used Edit Distances

- ED = WER (number of INS, DEL & SUB)
- ED_{swp} allow swap operator, i.e. `d(ab,ba) = d(ab,ab) = 0`
- ED_{sem} use semantic instead of morphologic information
- ED_{cnt} restrict comparison to content words, ignore functional words
- ED_{key} restrict comparison to keywords

RED (gRader based on Edit Distances)

Algorithm (learning)

1) Human labelling

   compute median score of human labels

2) Encode into 17-dimensional vector $M = M_1..M_{17}$

   M_1 = ED
   $M_2..M_{16}$ = all combinations of ED_{swp} ED_{sem} ED_{cnt} ED_{key}
   M_{17} = human score

3) Learn a decision tree with C4.5 algorithm

Algorithm (evaluation)

1) Redo step 2 w. $M_{17} = 0$ and apply learned decision tree to obtain M_{17}

RED (gRader based on Edit Distances)

Experiments

- comparison of 9 MT systems on sentence level and system level
- 9 human judges produced manual scores
- 10-fold cross validation

Method

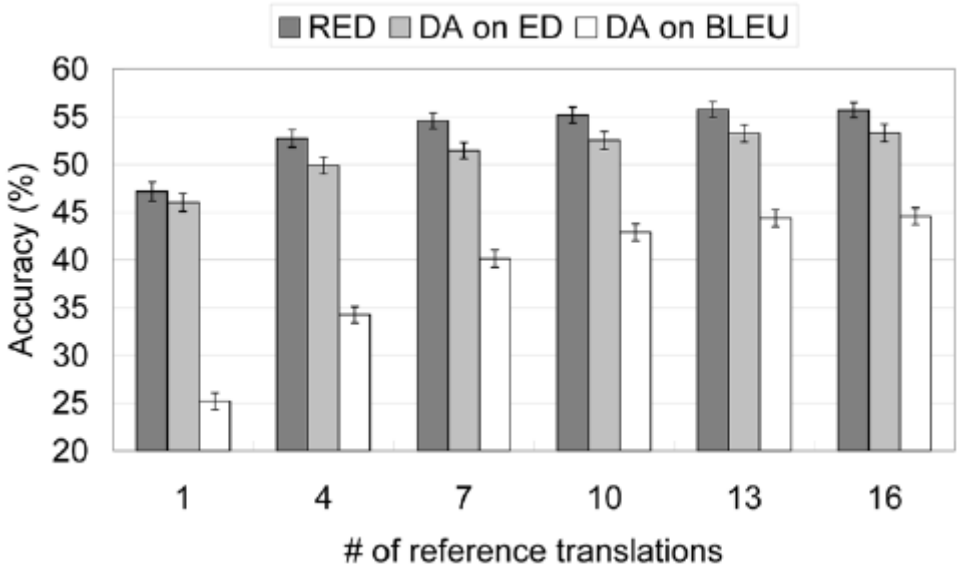
- sentence-level evaluation
  - discriminant analysis of scores for grades, accuracy measured
- system-level evaluation
  - statistical multiple comparison test of average sentence grades

Data

- 345 sentence pairs English – Japanese randomly chosen from BTEC corpus (topic: travelling, type: dialogues)
- 16 reference translations / sentence

RED (gRader based on Edit Distances)

Results



		RED					BLEU				
		# of reference translations					# of reference translations				
	XY	1	4	7	10	13	1	4	7	10	13
	00	21	20	19	20	20	0	0	0	0	0
	11	10	12	11	11	11	10	10	10	10	10
	22	2	2	2	2	2	1	1	1	1	1
Ratio (%)	00, 11 or 22	91.7	94.4	89.0	91.7	91.7	30.6	30.6	30.6	30.6	30.6

RED (gRader based on Edit Distances)

Conclusions

- RED outperforms BLEU on both, sentence level & system level comp.
  - higher agreement with human scores
- However:
  - simplified task (only 4 grades possible)
  - only shown for one language pair (English --> Japanese)
  - small evaluation corpus size

Minimum Error Rate Training

State-of-the-art: Training of statistical model parameters based on maximum likelihood et al. criteria

Problem: Difference in classification of error between statistical approach and automatic evaluation methods

- decision rule $\hat{\mathbf{e}}(\mathbf{f}) = \underset{\mathbf{e}}{\operatorname{argmax}} \{\Pr(\mathbf{e}|\mathbf{f})\}$ only optimal f. zero-one loss function
- other loss functions (e.g. BLEU) require different decision functions

Idea: Optimize model parameters with respect to evaluation criterion, e.g. BLEU, NIST, WER

Method: New training criterion f. log-linear MT model

Minimum Error Rate Training

Statistical MT with Log-linear models

- model posterior Pr(e|f) with *M* feature functions $h_m(e,f)$ with model parameters λ_m

$$\begin{aligned}\Pr(\mathbf{e}|\mathbf{f}) &= p_{\lambda_1^M}(\mathbf{e}|\mathbf{f}) \\ &= \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(\mathbf{e},\mathbf{f})]}{\sum_{e'_1{}^I} \exp[\sum_{m=1}^{M} \lambda_m h_m(e'_1{}^I,\mathbf{f})]}\end{aligned}$$

- Maximum mutual information criterion f. parameter optimization

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \left\{ \sum_{s=1}^{S} \log p_{\lambda_1^M}(\mathbf{e}_s|\mathbf{f}_s) \right\}$$

- Properties
  - unique global optimum
  - algorithms with guaranteed convergence (e.g. gradient descent)

Minimum Error Rate Training

New training criterion

- error counting function E(e,r) for sentence e against reference r
- candidate translations $C_s = \{e_{s,1},...,e_{s,K}\}$

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmin}} \left\{ \sum_{s=1}^{S} E(\mathbf{r}_s, \hat{\mathbf{e}}(\mathbf{f}_s; \lambda_1^M)) \right\} \quad (5)$$

$$= \underset{\lambda_1^M}{\operatorname{argmin}} \left\{ \sum_{s=1}^{S} \sum_{k=1}^{K} E(\mathbf{r}_s, \mathbf{e}_{s,k}) \delta(\hat{\mathbf{e}}(\mathbf{f}_s; \lambda_1^M), \mathbf{e}_{s,k}) \right\}$$

with

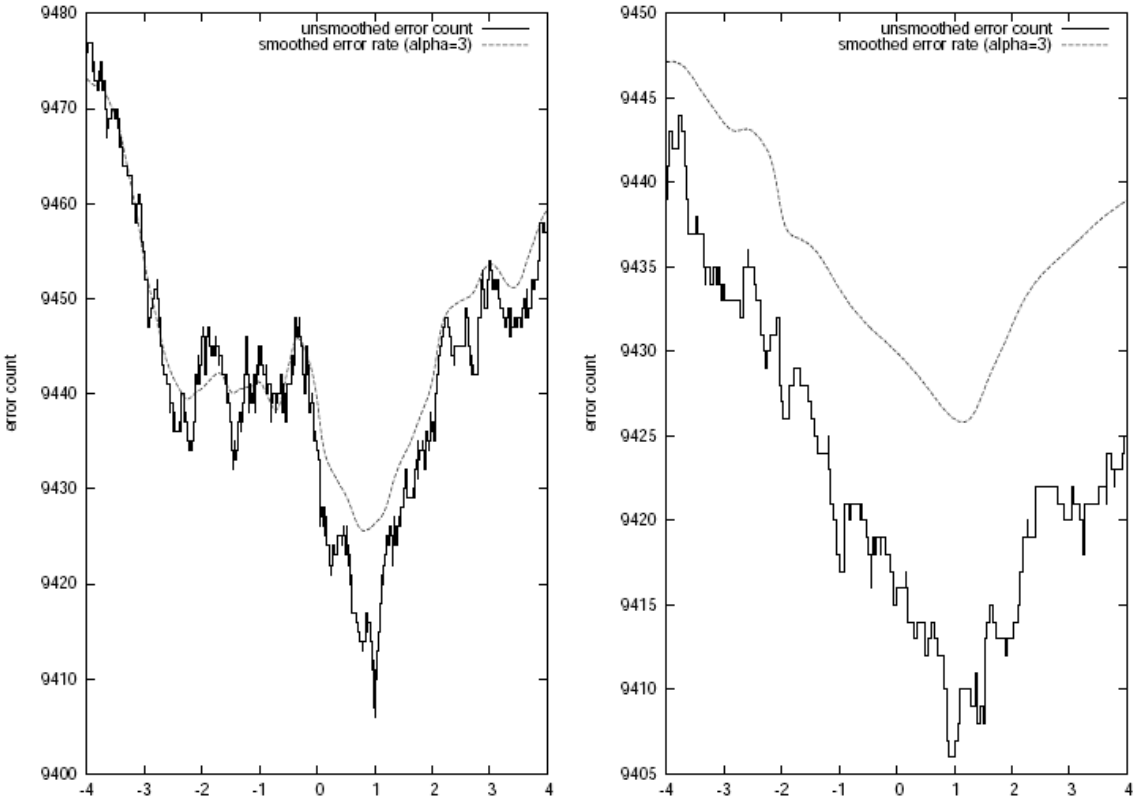
$$\hat{\mathbf{e}}(\mathbf{f}_s; \lambda_1^M) = \underset{\mathbf{e} \in \mathbf{C}_s}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}|\mathbf{f}_s) \right\} \quad (6)$$

- Problems
  - *argmax* prevents gradient descent
  - many local optima

Minimum Error Rate Training

Solution: Smoothing

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmin}} \left\{ \sum_{s,k} E(\mathbf{e}_{s,k}) \frac{p(\mathbf{e}_{s,k}|\mathbf{f})^\alpha}{\sum_k p(\mathbf{e}_{s,k}|\mathbf{f})^\alpha} \right\}$$



Minimum Error Rate Training

Optimization algorithm

- parameterize candidate translations in C as lines (t,m constant)

$$\hat{\mathbf{e}}(\hat{\mathbf{f}};\gamma) = \underset{\mathbf{e}\in\mathbf{C}}{\operatorname{argmin}} \{t(\mathbf{e},\mathbf{f}) + \gamma \cdot m(\mathbf{e},\mathbf{f})\}$$

- piecewise linear function

$$f(\gamma;\mathbf{f}) = \min_{\mathbf{e}\in\mathbf{C}} \{t(\mathbf{e},\mathbf{f}) + \gamma \cdot m(\mathbf{e},\mathbf{f})\}$$

- compute intervals $\gamma_1^{\mathbf{f}} < \gamma_2^{\mathbf{f}} < \ldots < \gamma_N^{\mathbf{f}}$ and incremental error count changes $\Delta E_1^{\mathbf{f}}, \Delta E_2^{\mathbf{f}}, \ldots, \Delta E_N^{\mathbf{f}}$ for each candidate sentence f ϵ C
- traverse sequence of interval boundaries & update error count to find minimum E
- update parameters according to interval for which min E was found

Minimum Error Rate Training

Experiments

- M=8 feature functions
  - e.g. language model logprob $h_1(e_1^I, f_1^J) = \log p_{\hat{\gamma}}(e_1^I)$

    translation model logprob $h_2(e_1^I, f_1^J) = \log p_{\hat{\theta}}(f_1^J | e_1^I)$
- dynamic programming beam search + *n*-best list from A* search
- pseudo-reference translations for MMI criterion
  - = sentences w. minimum word errors from n-best list

Data

- 2002 TIDES corpus, Chinese --> English

		Chinese	English
Train	Sentences	5 109	
	Words	89 121	111 251
	Singletons	3 419	4 130
	Vocabulary	8 088	8 807
Lex	Entries	82 103	
Dev	Sentences	640	
	Words	11 746	13 573
Test	Sentences	878	
	Words	24 323	26 489

Minimum Error Rate Training

Results

development set

error criterion used in training	mWER [%]	mPER [%]	BLEU [%]	NIST	# words
confidence intervals	+/- 2.4	+/- 1.8	+/- 1.2	+/- 0.2	-
MMI	70.7	55.3	12.2	5.12	10382
mWER	**69.7**	*52.9*	15.4	5.93	10914
smoothed-mWER	*69.8*	*53.0*	15.2	5.93	10925
mPER	*71.9*	**51.6**	17.2	6.61	11671
smoothed-mPER	*71.8*	*51.8*	17.0	6.56	11625
BLEU	76.8	54.6	**19.6**	*6.93*	13325
NIST	73.8	*52.8*	18.9	**7.08**	12722

test set

error criterion used in training	mWER [%]	mPER [%]	BLEU [%]	NIST	# words
confidence intervals	+/- 2.7	+/- 1.9	+/- 0.8	+/- 0.12	-
MMI	**68.0**	*51.0*	11.3	5.76	21933
mWER	*68.3*	*50.2*	13.5	6.28	22914
smoothed-mWER	*68.2*	*50.2*	13.2	6.27	22902
mPER	*70.2*	*49.8*	15.2	*6.71*	24399
smoothed-mPER	*70.0*	**49.7**	15.2	*6.69*	24198
BLEU	76.1	53.2	**17.2**	6.66	28002
NIST	73.3	*51.5*	*16.4*	**6.80**	26602

Minimum Error Rate Training

Conclusions

- Best performance for equal training error criterion / evaluation metric
- MMI is significantly worse except for mWER metric
- No difference between smoothed & unsmoothed error counts
  - small number of parameters
  - no overfitting

References

Y. Akiba, K. Imamura, E. Sumita, H. Nakaiwa, S. Yamamoto, H. G. Okuno, *Using, Multiple Edit Distances to Automatically Grade Outputs from Machine Translation Systems*. IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, No. 2, pp. 393-402, 2006.

Franz Josef Och, *Minimum Error Rate Training in Statistical Machine Translation*. In Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), 160-167, 2003.

Thank you