Structural Support Vector Machines for Log-linear Approach in Statistical Machine Translation

Katsuhiko Hayashi[†], Taro Watanabe[††], Hajime Tsukada[††], Hideki Isozaki[††]

†Doshisha University
††NTT Communication Science Laboratories

Outline

- Introduction
- Proposed Method
- Experiments
- Related Work
- Conclusion and Future Work

Introduction

- Background
  - Minimum Error Rate Training (MERT) is widely used.
- Problem
  - MERT tends to overfit to development data.
- Approach
  - We propose a training method that incorporates regularizer into objective function inspired by Structural Support Vector Machines.

Proposed Method (1/3)

- Objective Function inspired by 1-slack Structural SVM

Oracle Translation

$$\delta \mathbf{h}_s = \mathbf{h}_s(\mathbf{e}^*_s, \mathbf{f}_s) - \mathbf{h}_s(\hat{\mathbf{e}}_s, \mathbf{f}_s)$$

Feature vector of the reference

Feature vector of a translation

$$\min_{\mathbf{w}, \xi \geq 0} \frac{\lambda}{2} \|\mathbf{w}\|^2 + \xi$$

Model Parameter

Slack Variable

$$s.t.\ \forall (\hat{\mathbf{e}}_1, \cdots, \hat{\mathbf{e}}_S) \in \mathbf{C}^S : \frac{1}{S} \sum_{s=1}^{S} \langle \mathbf{w}, \delta \mathbf{h}_s \rangle \geq \Delta(\{\hat{\mathbf{e}}^*_s, \hat{\mathbf{e}}_s\}_1^S) - \xi$$

BLEU loss

$$\Delta(\{\hat{\mathbf{e}}^*_s, \hat{\mathbf{e}}_s\}_1^S) = Q \times \left(\mathrm{BLEU}(\{\mathbf{r}_s, \hat{\mathbf{e}}^*_s\}_1^S) - \mathrm{BLEU}(\{\mathbf{r}_s, \hat{\mathbf{e}}_s\}_1^S) \right)$$

Document-wise BLEU scores

Proposed Method (2/3)

- Och's Line Search Algorithm can be utilized for optimization.

Sco

$$\hat{\mathbf{e}}_{s,best} = \underset{\hat{\mathbf{e}}_s \in \mathbf{C}_s}{\operatorname{argmax}} \langle \mathbf{w} + \alpha \mathbf{d}, \mathbf{h}(\hat{\mathbf{e}}_s, \mathbf{f}_s) \rangle$$

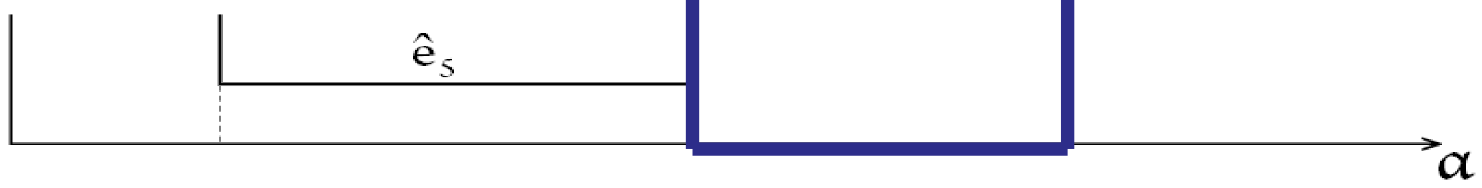
Update

$$\mathbf{w} + \hat{\alpha} \cdot \mathbf{d}$$

$$\hat{\alpha}$$

ble
sco:

$$= \underset{\hat{\mathbf{e}}_s \in \mathbf{C}_s}{\operatorname{argmax}} \left\{ \underbrace{\langle \mathbf{w}, \mathbf{h}(\hat{\mathbf{e}}_s, \mathbf{f}_s) \rangle}_{intercept} + \alpha \underbrace{\langle \mathbf{d}, \mathbf{h}(\hat{\mathbf{e}}_s, \mathbf{f}_s) \rangle}_{slope} \right\}$$



Proposed Method (3/3)

- The slope and intercept for each line
  - We need to calculate the slope and intercept by using the following equation

$$\underset{\hat{\mathbf{e}}_s \in \mathbf{C}_s}{\operatorname{argmax}} \left\{ \underbrace{\Delta(\mathbf{e}_s^*, \mathbf{e}_s) - \langle \mathbf{w}, \delta \mathbf{h}_s \rangle}_{intercept} + \alpha \underbrace{\langle \mathbf{d}, \delta \mathbf{h}_s \rangle}_{slope} \right\}$$

However, the slope and intercept calculated by this equation are very noisy because of sentence-wise BLEU scores.

Comparison to MERT

	Objective Function	Regularizer	Hyper Paramter	Optimization
MERT	Evaluation Metrics (BLEU)	×	×	Line Search Algorithm
Proposed Method	L2 Regularizer + Emprical Risk (incorpolated with BLEU loss)	○	○	Line Search Algorithm, SVM^{struct}

Advantage

- Our proposed method is a natural extension to regularize MERT's objective function.
- It is easy to implement
  - We can use almost the same line search algorithm as used in MERT.

Experiments

- **Goal**
  - **To investigate a validity of our proposed method, compared with MERT.**
  - **Compare generalization ability**
    - In case of out-of-domain
    - With sparse data

Common Settings

- **Decoder**
- Moses (Koehn et al., 2007)
  - 14 real-valued features

- **Translation Model**
- GIZA++ (Och et al., 2003)

- **Language Model**
- SRILM (et al., 2002)

Data Set

- Europarl French-English WMT08-shared task
- **Training data**
  - 1.28M (Europarl)
- **Development data**
  - 2.0K (Europarl)
- **Test data**
  - In-domain test set 2.0k (Europarl)
  - out-of-domain test set 1.5k (News)

Hyper Parameters

- Two hyper parameters were tuned by Cross Validation Method.

$$\min_{\mathbf{w}, \xi \geq 0} \frac{\lambda}{2} \|\mathbf{w}\|^2 + \xi$$

$$s.t. \ \forall (\hat{\mathbf{e}}_1, \cdots, \hat{\mathbf{e}}_S) \in \mathbf{C}^S : \quad \frac{1}{S} \sum_{s=1}^{S} \langle \mathbf{w}, \delta \mathbf{h}_s \rangle \geq \Delta(\{\hat{\mathbf{e}}_s^*, \hat{\mathbf{e}}_s\}_1^S) - \xi$$

$$\Delta(\{\hat{\mathbf{e}}_s^*, \hat{\mathbf{e}}_s\}_1^S) = Q \times \left\{ \mathbf{BLEU}(\{\mathbf{r}_s, \hat{\mathbf{e}}_s^*\}_1^S) - \mathbf{BLEU}(\{\mathbf{r}_s, \hat{\mathbf{e}}_s\}_1^S) \right\}$$

λ emphasizes the convex regularizer.

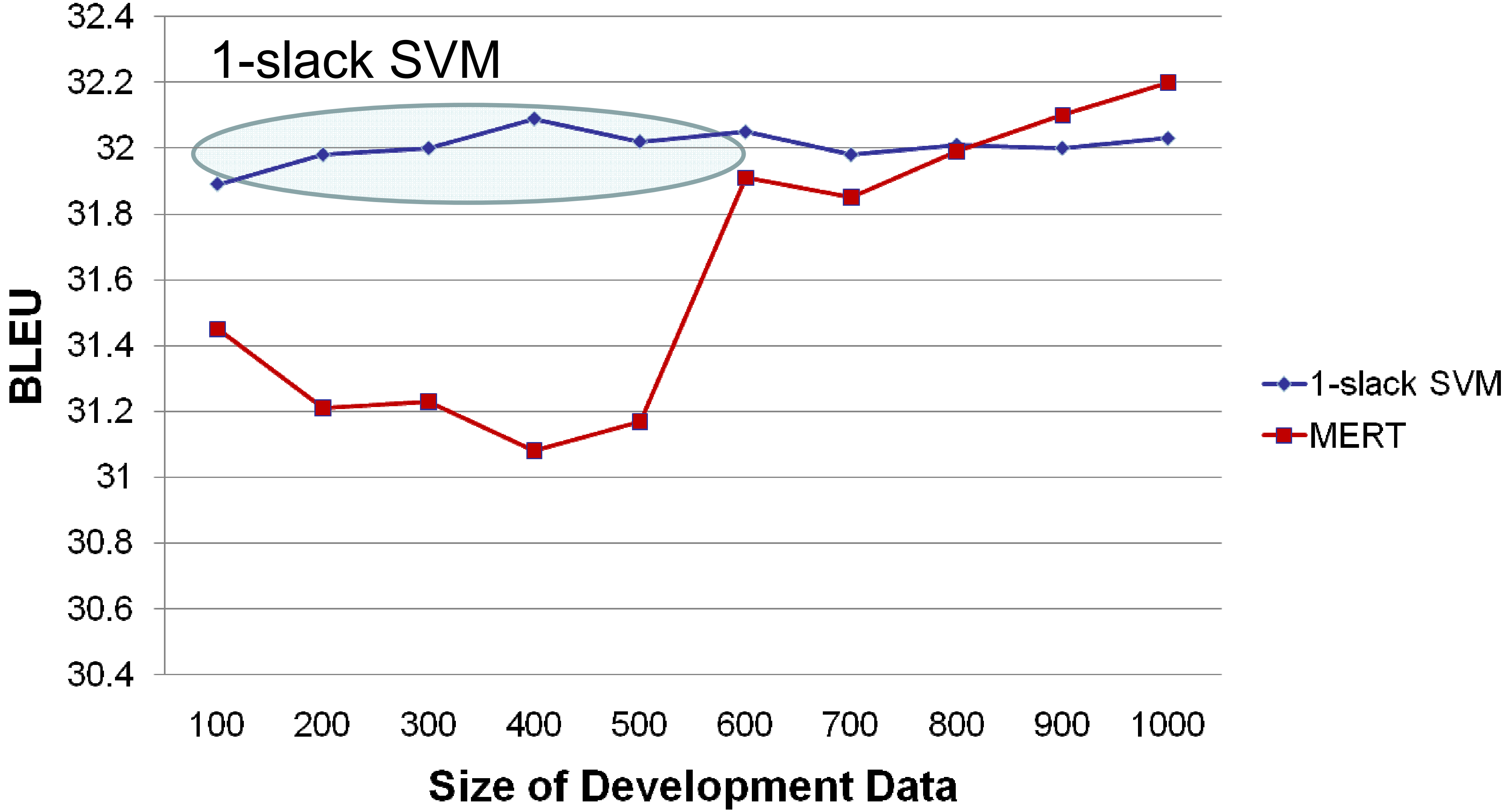
Q is a constant for scaling the BLEU scores.

Result

- In-Domain vs Out-of-Domain -

	In-Domain	Out-of-Domain
MERT	32.36	13.81
Smoothed-MERT (Och 03)	31.96	13.76
Proposed Method	32.42	14.13

Size of Development data



Data Sparseness

- We reduced development data (2.0K).
  - 400 sentences randomly selected from a full development data
  - Experiments were conducted 4 times

- We expected our proposed method to reduce overfitting problem.

Result
- Data Sparsness -

- The average BLEU scores on 4 times experiments

	In-Domain
MERT	31.24
Smoothed MERT (Och, 03)	31.06 (-0.18)
Proposed Method	31.76 (+0.52)

Related Work

- Och (2003) tried to regularize MERT's objective function by using the same regularization as used in Speech Community.
- Cer (2008) proposed window smoothing method with line search algorithm.
- Watanabe (2006) applied Margin Infused Relaxed Algorithm to Statistical Machine Translation.

Conclusion

- We proposed a learning method for SMT by using a objective function inspired by Structural SVM.
- The objective function involves both document-wise BLEU and a regularizer.
- The proposed method (1-slack) outperforms MERT when the development data size is small.

Conclusion (2/2)

◆Future Work

- We will apply 1-slack SVM to the decoder which has a large number of features.

- In this case, SVM*struct* may be a more appropriate optimization algorithm.

Thank you very much for your attention !!