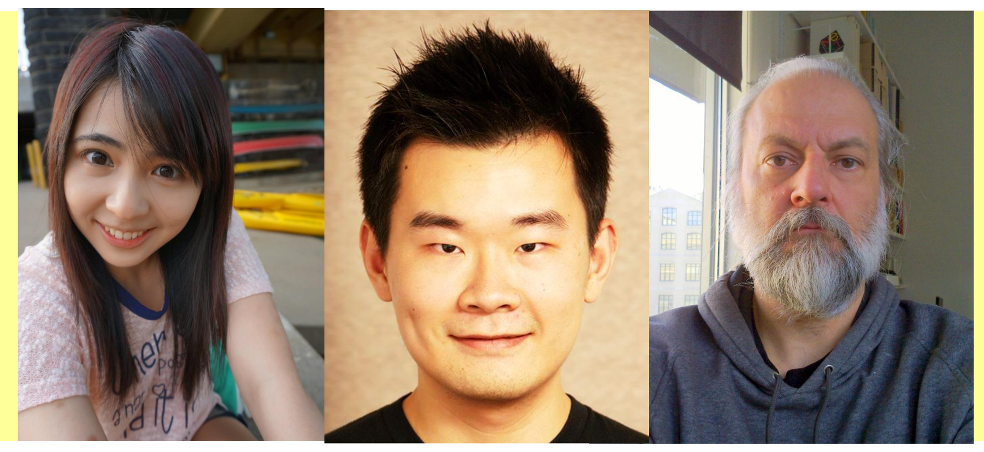
Carnegie Mellon

An Empirical Investigation of Sparse Log-Linear Models for Improved Dialogue Act Classification

Yun-Nung (Vivian) Chen, William Yang Wang, and Alexander I. Rudnicky



1. Summary

- Motivations
  - ASR outputs are often noisy
  - Dense models might overfit to the training data
  - Sparse models maintain a compact feature space, which is robust to noise
- Approaches
  - Element-wise sparsity: lasso, ridge, elastic net
  - Structured sparsity
  - Hierarchical sparsity
- Results
  - 19.7% improvement over a rule-based baseline
  - 3.7% improvement over a traditional non-sparse log-linear model
  - outperformed a state-of-the-art SVM model by 2.2%

2. The Materials

- The corpus
  - Domain: restaurant recommendation in Cambridge [1] (WER = 37%)
  - Dialogue act (total #act = 17):
    inform, request, bye, null, affirm, hello, negate, reqalts, confirm, thankyou, others (< 0.8%)
- Feature set (N = 10)
  - W_1: word trigram freq. from 1-best hypothesis
  - W_N: word trigram freq. from N-best hypothesis
  - P_1: phone trigram freq. from 1-best hypothesis
  - P_N: phone trigram freq. from N-best hypothesis
  - CNet: word confusion networks with context freq.

	Training	Testing
Dialogues	1522	644
Utterances	10571	4882
Male:Female	28:31	15:15
Native:Non-Native	33:26	21:9

3. Log-Linear Models

- Multinomial logistic regression (MLR)
  - Multiclass classification $\hat{y} \sim \mathrm{Mult}(\hat{\theta})$
  - K instances, M classes

$$\hat{\theta}_{im} = \frac{\exp(Z_{mi})}{\sum_{m=1}^{M}\exp(Z_{mi})} \qquad Z_{mi} = c_m + \sum_{d=1}^{D} \boxed{\beta_{md}}\boxed{X_{id}}$$

the d-th feature of instance i

puts a weight on feature X_d for predicting the class label

$$\ell(\theta) = \sum_{i=1}^{K}\sum_{m=1}^{M} y_{im}\log\theta_{im}$$

  - using the standard maximum likelihood estimation approach, the parameters β_{md} can be set by the gradient ascent approach
  - using the L-BFGS implementation for the numerical optimization of sparse models

- Element-wise sparsity
  - Lasso

$$\min\Big(-\ell(\theta) + \boxed{\sum_{m=1}^{M}\sum_{d=1}^{D}\lambda_m^{(1)}\|\beta_{md}\|}\Big)$$

L_1-norm

➢ discontinuities to the original convex function

  - Ridge

$$\min\Big(-\ell(\theta) + \boxed{\sum_{m=1}^{M}\sum_{d=1}^{D}\lambda_m^{(2)}\|\beta_{md}\|^2}\Big)$$

L_2-norm

➢ quadratic penalty maintains the convex property

  - Elastic net

$$\min\Big(-\ell(\theta) + \boxed{\sum_{m=1}^{M}\sum_{d=1}^{D}\lambda_m^{(1)}\|\beta_{md}\| + \sum_{m=1}^{M}\sum_{d=1}^{D}\lambda_m^{(2)}\|\beta_{md}\|^2}\Big)$$

L_1+L_2-norm

➢ balances the sparsity and smoothness properties

- Structured sparsity

group lasso

$$\min\Big(-\ell(\theta) + \boxed{\sum_{m=1}^{M}\sum_{g=1}^{G}\lambda_m\|\beta_{gm}\|}\Big)$$

➢ modeling the dependency and interaction of groups of local features

$$\min\Big(-\ell(\theta) + \boxed{\sum_{m=1}^{M}\max_{d}\lambda_m^{(1)}\|\beta_{md}\|}\Big)$$

➢ reveals the important features across different output classes

$L_{1,\mathrm{inf}}$-norm

- Hierarchical sparsity

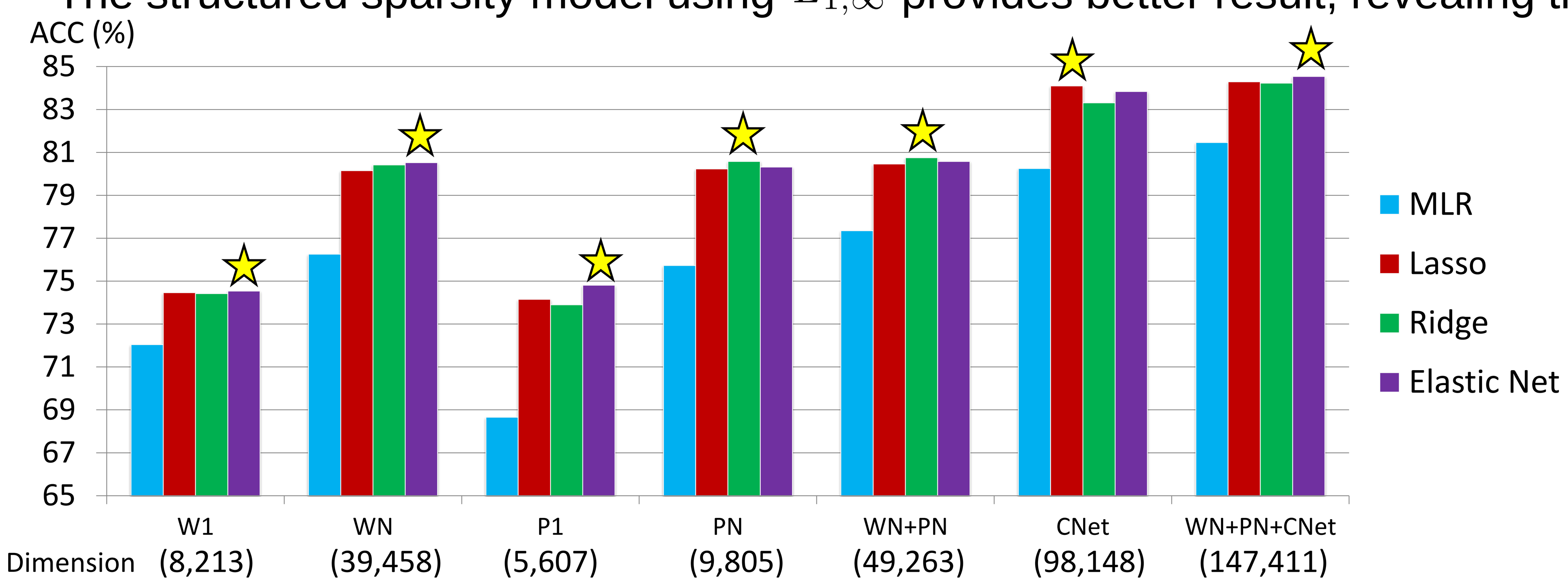
$$\min\Big(-\ell(\theta) + \boxed{\sum_{m=1}^{M}\sum_{g=1}^{G}\lambda_m\|\beta_{gm}\| + \sum_{m=1}^{M}\sum_{d=1}^{D}\lambda_m^{(1)}\|\beta_{md}\|}\Big)$$

L_1–norm + group lasso

➢ combines the element-wise and the group-wise lasso

4. Empirical Evaluation

- The improvement of sparse models over MLR with W_N/P_N is greater than with W_1/P_1, because using N-best hypotheses allows the sparse models to make use of more information.
- Both W_N and P_N features have obtained significant improvements over MLR baseline when using sparse models, demonstrating the robustness of our sparse models to filter noisy features in the settings with distinct dimensionalities.
- Combining three feature sets can further improve the performance.
- Elastic net model that balances sparsity and smoothness obtains the best performance.
- The structured sparsity model using $L_{1,\infty}$ provides better result, revealing the importance of modeling sparsity structures.



Model		ACC (%)
Element-wise	Lasso	84.29
Structured	Group Lasso	83.39
	$L_{1,\infty}$	**84.41**
Hierarchical	Sparse Group Lasso	83.35

Model	Feature	ACC (%)
Phoenix	manual grammar	70.6 ± 1.28
SVM	W_N+P_N+CNet	82.7 ± 1.06
MLR		81.5 ± 1.09
Best Sparse MLR		**84.5 ± 1.02**

5. Conclusions

- Sparse log-linear models improve dialogue act classification
  - absolute improvements over several baselines and a state-of-the-art SVM model (from 2.2% to 19.7%)
  - the improvements are robust across different features and parameter settings
- Sparse models have larger gains on the word-level N-best ASR hypotheses than that on the 1-best hypothesis
- Augmenting the word-level n-gram and confusion network features with phonetic features in our sparse models performs best.
- Empirical results show that the elastic net model that balances sparsity and smoothness obtains the best overall performance
- The $L_{1,\infty}$ structured sparsity model yields promising results among structured and hierarchical sparse models.

[1] M. Henderson, M. Gaˇsi´c, B. Thomson, P. Tsiakoulis, K. Yu, and S. Young, "Discriminative spoken language understanding using word confusion networks," in SLT, 2012.