		F1-Score	
Embedding/Model		Dev. Set	Test Set
Baseline	No Gazetteers	90.03	84.39
C&W, 200-dim		92.46	87.46
HLBL, 100-dim		92.00	88.13
Brown 1000 clusters		92.32	88.52
Ando & Zhang '05		93.15	89.31
Suzuki & Isozaki '08		93.66	89.36
LR-MVL (CO) $50 \times$ 3-dim		93.11	89.55
LR-MVL $50 \times$ 3-dim		93.61	**89.91**
HLBL, 100-dim	With Gazetteers	92.91	89.35
C&W, 200-dim		92.98	88.88
Brown, 1000 clusters		93.25	89.41
LR-MVL (CO) $50 \times$ 3-dim		93.91	89.89
LR-MVL $50 \times$ 3-dim		94.41	**90.06**

Table 1: NER Results. **Note:** 1). LR-MVL (CO) are Context Oblivious embeddings which are gotten from (**A**) in Algorithm 1. 2). F1-score= Harmonic Mean of Precision and Recall. 3). The current state-of-the-art for this NER task is 90.90 (Test Set) but using 700 billion tokens of unlabeled data [19].

Embedding/Model	Test Set F1-Score
Baseline	93.79
HLBL, 50-dim	94.00
C&W, 50-dim	94.10
Brown 3200 Clusters	94.11
Ando & Zhang '05	94.39
Suzuki & Isozaki '08	94.67
LR-MVL (CO) $50 \times$ 3-dim	95.02
LR-MVL $50 \times$ 3-dim	**95.44**

Table 2: Chunking Results.

It is important to note that in problems like NER, the final accuracy depends on performance on rare-words and since LR-MVL is robustly able to correlate past with future views, it is able to learn better representations for rare words resulting in overall better accuracy. On rare-words (occurring < 10 times in corpus), we got 11.7%, 10.7% and 9.6% relative reduction in error over C&W, HLBL and Brown respectively for NER; on chunking the corresponding numbers were 6.7%, 7.1% and 8.7%.

Also, it is worth mentioning that modeling the context in embeddings gives decent improvements in accuracies on both NER and Chunking problems. For the case of NER, the polysemous words were mostly like *Chicago, Wales, Oakland etc.*, which could either be a *location* or *organization* (Sports teams, Banks etc.), so when we don't use the gazetteer features, (which are known lists of cities, persons, organizations etc.) we got higher increase in F-score by modeling context, compared to the case when we already had gazetteer features which captured most of the information about polysemous words for NER dataset and modeling the context didn't help as much. The polysemous words for Chunking dataset were like *spot (VP/NP), never (VP/ADVP), more (NP/VP/ADVP/ADJP) etc.* and in this case embeddings with context helped significantly, giving $3.1 - 6.5\%$ relative improvement in accuracy over context oblivious embeddings.

5 Summary and Conclusion

In this paper, we presented a novel CCA-based multi-view learning method, LR-MVL, for large scale sequence learning problems such as arise in NLP. LR-MVL is a spectral method that works in low dimensional state-space so it is computationally efficient, and can be used to train using large amounts of unlabeled data; moreover it does not get stuck in local optima like an EM trained HMM. The embeddings learnt using LR-MVL can be used as features with any supervised learner. LR-MVL has strong theoretical grounding; is much simpler and faster than competing methods and achieves state-of-the-art accuracies on NER and Chunking problems.

Acknowledgements: The authors would like to thank Alexander Yates, Ted Sandler and the three anonymous reviews for providing valuable feedback. We would also like to thank Lev Ratinov and Joseph Turian for answering our questions regarding their paper [16].