

AN EXPERIMENTAL COMPARISON OF ACTIVE LEARNING STRATEGIES FOR PARTIALLY LABELED SEQUENCES

Diego Marcheggiani[†], Thierry Artières[§]

[†] ISTI - Consiglio Nazionale delle Ricerche, Pisa, Italy,
diego.marcheggiani@isti.cnr.it

[§] LIP6, Pierre et Marie Curie University, Paris, France,
thierry.artieres@lip6.fr

We experimentally investigate the behavior of several AL strategies for sequence labeling tasks (in a partially-labeled scenario) tailored on Partially-Labeled Conditional Random Fields, on four sequence labeling tasks: phrase chunking, part-of-speech tagging, named-entity recognition, and bio-entity recognition.

INTRODUCTION

One of the main problem of machine learning approaches lies in their need of large human-annotated training data. The process of *active learning* (AL) asks human annotators to label new samples which are supposed to be the most informative in the creation of a new classifier. In this work we focus on AL strategies for partially labeled sequences adopting the single token, instead of the entire sequence, as annotation unit and Partially-Labeled CRFs (PL-CRFs) [2] as learning algorithm.

Partially-Labeled CRFs

PL-CRFs allow to learn a CRF model using partially-labeled sequences.

The marginal probability $p(y_t = j | \mathbf{x}, \mathbf{L})$ is calculated as: $p(y_t = j | \mathbf{x}, \mathbf{L}) = \frac{\alpha_{t,L}(j) \cdot \beta_{t,L}(j)}{Z_L(\mathbf{x})}$, where \mathbf{L} denotes a partially labeled information about a sequence.

The most probable sequence assignment \mathbf{y}^* is calculated by the Viterbi algorithm.

ACTIVE LEARNING STRATEGIES

Algorithm 1 Pool-based active learning framework

Require: the training set \mathcal{T}_1 , the unlabeled set \mathcal{U}_1 , the AL strategy \mathcal{S} , the number of iterations n , the dimension of the update batch B

```

1: for  $i \leftarrow 1$  to  $n$  do
2:    $\Phi_i \leftarrow \text{train}(\mathcal{T}_i)$ 
3:    $\mathcal{L}_i \leftarrow \Phi_i(\mathcal{U}_i)$ 
4:   for  $b \leftarrow 1$  to  $B$  do
5:      $\mathbf{x}_*^{(b)} \leftarrow \arg \min_{\mathbf{x}_t \in \mathcal{L}_i, \mathbf{x} \in \mathcal{L}_i} \mathcal{S}(t, \mathbf{x})$ 
6:      $\mathcal{L}_i \leftarrow \mathcal{L}_i - \mathbf{x}_*^{(b)} \cup \Phi_i(\mathbf{x}_*^{(b)}, y_*)$  // re-estimation step
7:      $\mathcal{U}_i \leftarrow \mathcal{U}_i - \mathbf{x}_*^{(b)} \cup (\mathbf{x}_*^{(b)}, y_*)$ 
8:      $\mathcal{T}_i \leftarrow \mathcal{T}_i - \mathbf{x}_*^{(b)} \cup (\mathbf{x}_*^{(b)}, y_*)$ 
9:    $\mathcal{U}_{i+1} \leftarrow \mathcal{U}_i$ 
10:   $\mathcal{T}_{i+1} \leftarrow \mathcal{T}_i$ 

```

- For each iteration through the update batch B , the most informative element $\mathbf{x}_*^{(b)}$, according to the AL strategy \mathcal{S} , is chosen.
- \mathcal{L}_i is the set that contains the tokens automatically labeled by the classifier Φ_i and the information (e.g., confidence) associated to them.
- After the choice of the most informative token, the sets \mathcal{L} , \mathcal{U} , and \mathcal{T} are updated.

Greedy Strategies

Select the most informative tokens regardless of the assignment performed by the Viterbi algorithm.

- Minimum Token Probability (MTP)

$$\mathcal{S}^{MTP}(t, \mathbf{x}) = \max_{j \in \mathcal{Y}} p(y_t = j | \mathbf{x}, \mathbf{L})$$

- Maximum Token Entropy (MTE)

$$\mathcal{S}^{MTE}(t, \mathbf{x}) = \sum_{j \in \mathcal{Y}} p(y_t = j | \mathbf{x}, \mathbf{L}) \cdot \log p(y_t = j | \mathbf{x}, \mathbf{L})$$

- Minimum Token Margin (MTM)

$$\mathcal{S}^{MTM}(t, \mathbf{x}) = \max_{j \in \mathcal{Y}} p(y_t = j | \mathbf{x}, \mathbf{L}) - \max_{j' \in \mathcal{Y}'} p(y_t = j' | \mathbf{x}, \mathbf{L})$$

Viterbi Strategies

The most informative tokens are chosen according to the information obtained from the outcome of the Viterbi algorithm (i.e., the most probable sequence assignment).

- Minimum Viterbi Probability (MVP) introduced by [3]

$$\mathcal{S}^{MVP}(t, \mathbf{x}) = p(\mathbf{y}_t^* | \mathbf{x}, \mathbf{L})$$

- Maximum Viterbi Pseudo-Entropy (MVPE)

$$\mathcal{S}^{MVPE}(t, \mathbf{x}) = \sum_{j \in \mathcal{Y}} p(\mathbf{y}_{y_t=j}^* | \mathbf{x}, \mathbf{L}) \cdot \log p(\mathbf{y}_{y_t=j}^* | \mathbf{x}, \mathbf{L})$$

- Minimum Viterbi Margin (MVM)

$$\mathcal{S}^{MVM}(t, \mathbf{x}) = p(\mathbf{y}_{y_t}^* | \mathbf{x}, \mathbf{L}) - p(\mathbf{y}_{y_t}^{\prime*} | \mathbf{x}, \mathbf{L})$$

- Minimum Expectation (ME)

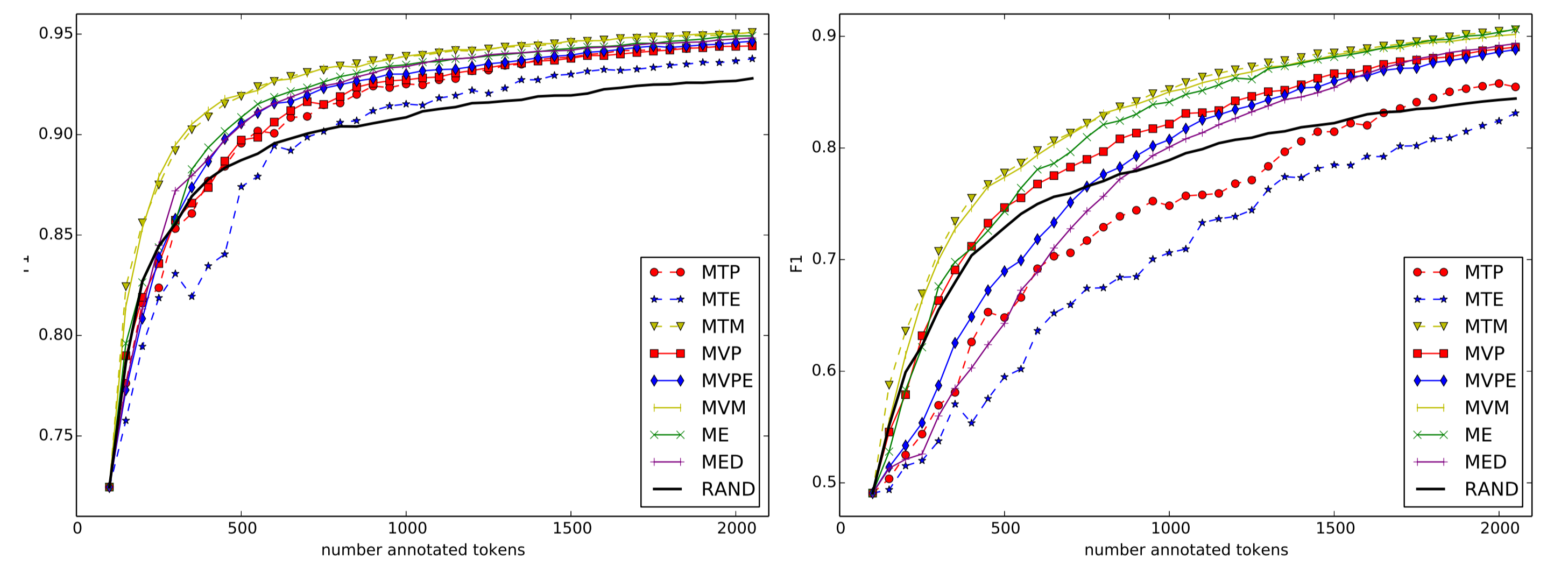
$$\mathcal{S}^{ME}(t, \mathbf{x}) = \sum_{j \in \mathcal{Y}} p(y_t = j | \mathbf{x}, \mathbf{L}) \cdot p(\mathbf{y}_{y_t=j}^* | \mathbf{x}, \mathbf{L})$$

- Minimum Expectation Difference (MED)

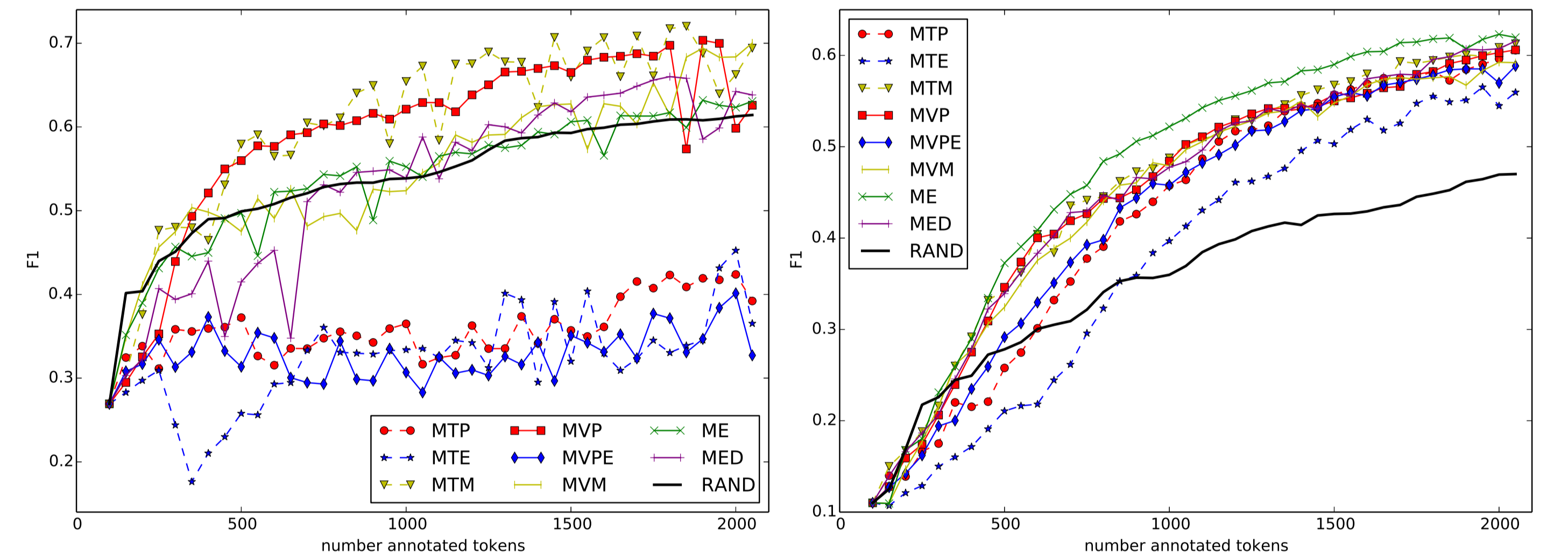
$$\mathcal{S}^{MED}(t, \mathbf{x}) = \mathcal{S}^{ME}(t, \mathbf{x}) - p(\mathbf{y}^* | \mathbf{x}, \mathbf{L})$$

The Random (RAND) strategy samples random tokens without any external information.

RESULTS



F₁ results on chunking task (left) and POS tagging task (right).



F₁ results on NER task (left) and Bio-NER task (right).

- MTE and MTP perform particularly bad in all the tasks.
- MTM is very effective in most of the tasks.
- MTM, MVM, and MVP, perform very good in all the tasks.
- ME strategy is always above the average and it is the best strategy in the Bio-NER task.
- The AL strategies applied on the NER task suffer of some “random” drop of performance.
- This phenomenon is probably due to the *missed class effect* [1].

CONCLUSIONS

We have presented several AL strategies tailored for PL-CRFs in a pool-based scenario. We have tested the proposed strategies on four different datasets for four different sequence labeling tasks. Differently from other similar work in the field of AL, in this study we have shown that margin-based strategies constantly achieve good performance on four tasks with very different data characteristics.

REFERENCES

- [1] K. Tomanek, F. Laws, U. Hahn, and H. Schütze. On proper unit selection in active learning: co-selection effects for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 9–17, Boulder, US, 2009.
- [2] Y. Tsuboi, H. Kashima, H. Oda, S. Mori, and Y. Matsumoto. Training conditional random fields using incomplete annotations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 897–904, Manchester, UK, 2008.
- [3] D. Wanvarie, H. Takamura, and M. Okumura. Active learning with subsequence sampling strategy for sequence labeling tasks. *Information and Media Technologies*, 6(3):680–700, 2011.