Fraunhofer
SCAI

FRAUNHOFER INSTITUTE FOR ALGORITHMS AND SCIENTIFIC COMPUTING SCAI

USER'S CHOICE OF PRECISION AND RECALL IN NAMED ENTITY RECOGNITION

Roman Klinger and Christoph M. Friedrich

Problem Description

- Applications have different demands on Named Entity Recognition, e. g.
  - Information Extraction: High precision is needed to only extract true information
  - Information Retrieval: High recall to not miss important resources
- ⇒ Decision can only be made by the user, not by the developer!

Example data to find person names:

Text:	"… John F. Kennedy inspired the name of the airport JFK …"		
Annotation:	John F. Kennedy		⇐ perfect, but difficult
Annotation:			⇐ high precision
Annotation:	John F. Kennedy	JFK	⇐ high recall

- Precision $\text{prec} = \frac{TP}{TP+FP}$ • Recall $\text{rec} = \frac{TP}{TP+FN}$ • $F_\beta = \frac{(1+\beta^2)\cdot\text{prec}\cdot\text{rec}}{\beta^2\cdot\text{prec}+\text{rec}}$

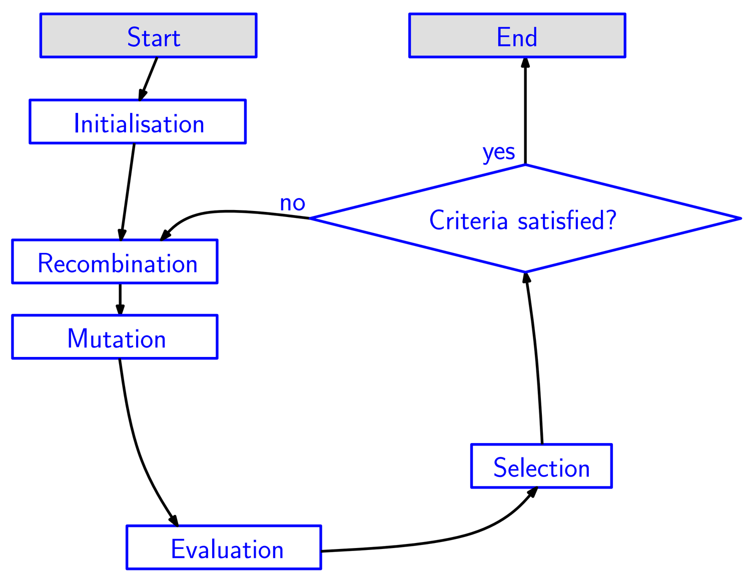
Background

- Conditional Random Fields are a class of probabilistic graphical models
  - Typical Application in NLP: Text Segmentation, e.g. Named Entity Recognition
- Conditional Random Fields are typically trained to maximize accuracy (via maximum log-likelihood and gradient-based optimization)
- Evaluation is typically performed wrt. F_1 measure
- ↯ What is really what we need?
  - ⇒ Depends on individual requirements!

$\max F_{0.5}$ ⇒ High Precision, $\max F_2$ ⇒ High Recall, $\max F_1$ ⇒ Similar Precision and Recall

Existing Solutions

- Train what you need [4]
  - Smooth objective function, train gradient-based
  - → Application needs to be known at training time: not always possible
- Select solutions at inference step [1]
  - Train via maximum likelihood, compute confidences of solutions with forward-backward algorithm, set threshold
  - → Decreases Speed



Multi Objective Optimization

- Non-Dominated Sorting Genetic Algorithm II (NSGA-II) optimizes multiple objective functions
- Pareto-Optimal set of solutions of non-dominated solutions is provided
  - Non-Domination: No solutions exist with at least one better objective value
- Evolutionary Algorithm: Specification needed for
  - Initialization
  - Variation: Mutation, Recombination
  - Selection via objective functions

MOCRF

Multi Objective Optimization of CRF:

Initialization Initial parameters found by optimizing parameters $\vec{\lambda}$ wrt. log-likelihood $\log P_{\vec{\lambda}}(\vec{y}|\vec{x})$ (with token sequence $\vec{x}$ and segmentation sequence $\vec{y}$)

Mutation $\forall k : \text{mut}(\lambda_k) = \lambda_k + \mathcal{N}(0, \sigma)$ with stepsize $\sigma = 0.01$

Recombination Select randomly from

$$\text{im}(\vec{\lambda}_1, \vec{\lambda}_2) = \left((\lambda_{1,1} + \lambda_{2,1})/2, \dots, (\lambda_{1,n} + \lambda_{2,n})/2\right)^T$$

$$\text{co}(\vec{\lambda}_1, \vec{\lambda}_2) = \left(\lambda_{1,1}, \dots, \lambda_{1,r}, \lambda_{2,r+1}, \dots, \lambda_{2,n}\right)^T$$

Objective Functions

- Precision • Recall

Results

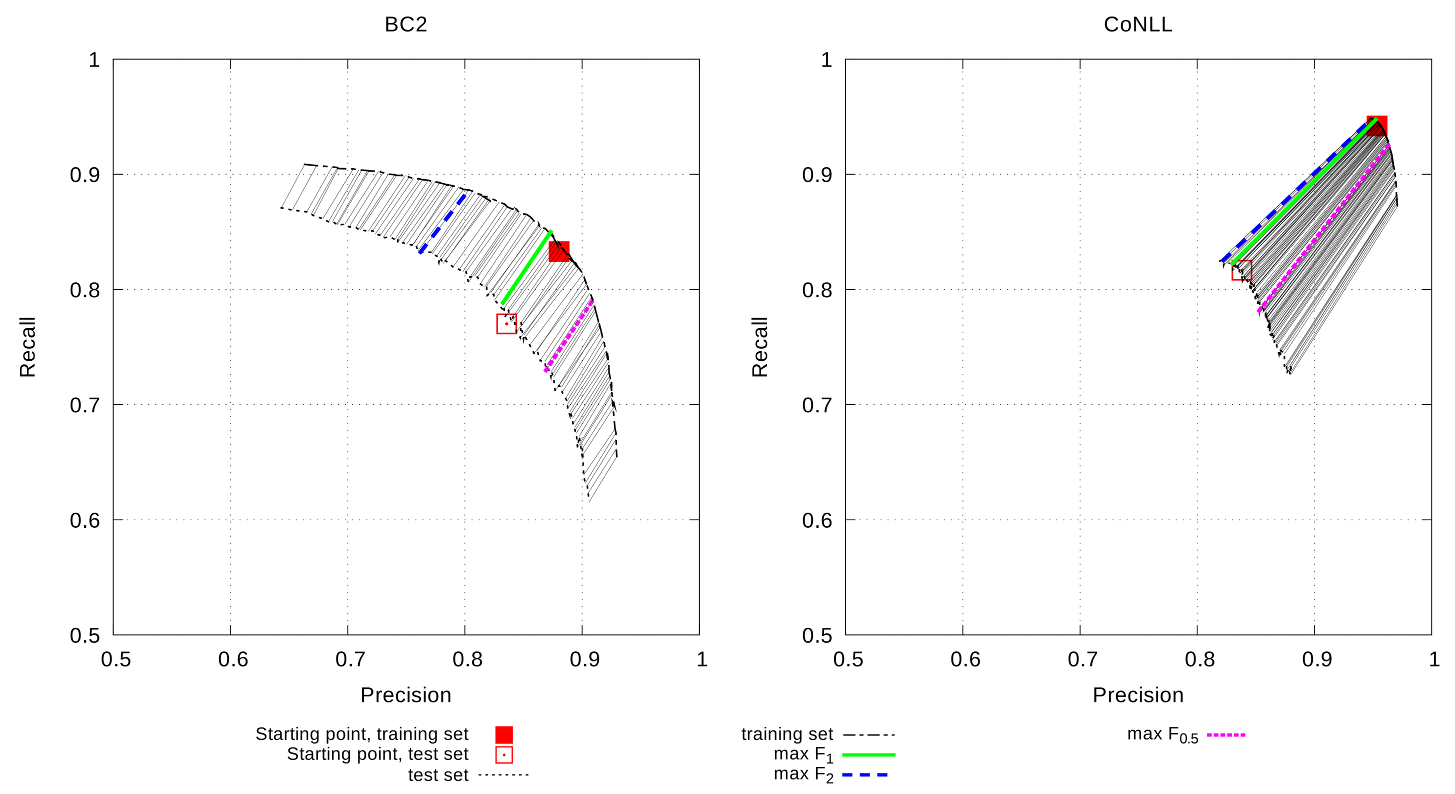
Setting

- Data Sets
  - BioCreative 2: Gene and Protein Names
  - CoNLL 2003: Persons, Organizations, Locations, Misc.
- CRF with fairly standard feature set, feature selection [2]
  - 23000 features and 38000 features
- 100 individuals, 100 iterations for genetic algorithm

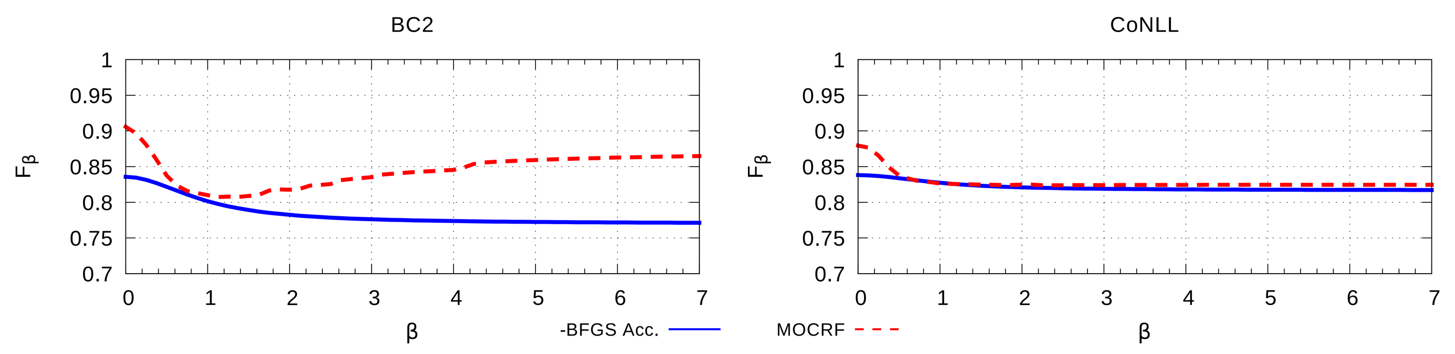
Discussion

- Method provides set of solutions with different precision/recall
- User can select the appropriate model for the particular application without additional computational complexity during application
- Working well for BioCreative, small increase of recall for CoNLL
  - → Problem of multiple entities of interest
- Result in F_β measure greater than for classical likelihood optimization for nearly all $\beta \in [0, 7]$

Training pareto-front and independent test



Best F_β, multiple solutions vs. one likelihood-based solution



References

[1] B. Carpenter. LingPipe for 99.99 % Recall of Gene Mentions. In *Proceedings of the 2nd BioCreative workshop*, Madrid, Spain, 2007.

[2] R. Klinger and C. M. Friedrich. Feature Subset Selection in Conditional Random Fields for Named Entity Recognition. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria, 2009.

[3] R. Klinger and C. M. Friedrich. User's Choice of Precision and Recall in Named Entity Recognition. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria, 2009.

[4] J. Suzuki, E. McDermott, and H. Isozaki. Training Conditional Random Fields with Multivariate Evaluation Measures. In *Proceedings of the ACL*, pages 217–224, 2006.

This work is partially funded by the MPG-FHG Machine Learning Collaboration: http://lip.fml.tuebingen.mpg.de/

Contact

Roman Klinger
roman.klinger@scai.fraunhofer.de
http://www.scai.fraunhofer.de