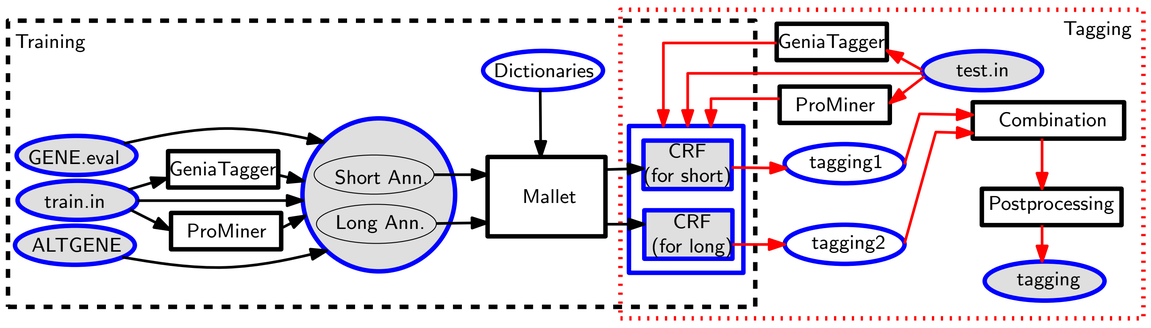
Multi Model Approach for Alternative Taggings

Roman Klinger, Christoph M. Friedrich and Juliane Fluck
Department of Bioinformatics

Fraunhofer Institute Algorithms and Scientific Computing

Overview

1. Generate Trainingdata depending on different lengths of annotations
2. Build different Conditional Random Fields
3. Tag testdata and combine it with different strategies
4. Postprocessing
   - Bracket Correction, Acronym Disambiguation using LSA



Problem Description

- Characteristic in BioCreative 2006:
  - Trainingdata provides acceptable alternatives additional to gold standard
- Problem: Ambiguities — Examples:
  - On the other hand factor IX activity is decreased in coumarin treatment with factor IX antigen remaining normal.
  - The arginyl peptide bonds that are cleaved in the conversion of human factor IX to factor IXa by factor XIa were identified as Arg145-Ala146 and Arg180-Val181.

  (Gold Standard Alternative)

Multi Model Approach

How to use the alternative annotations?

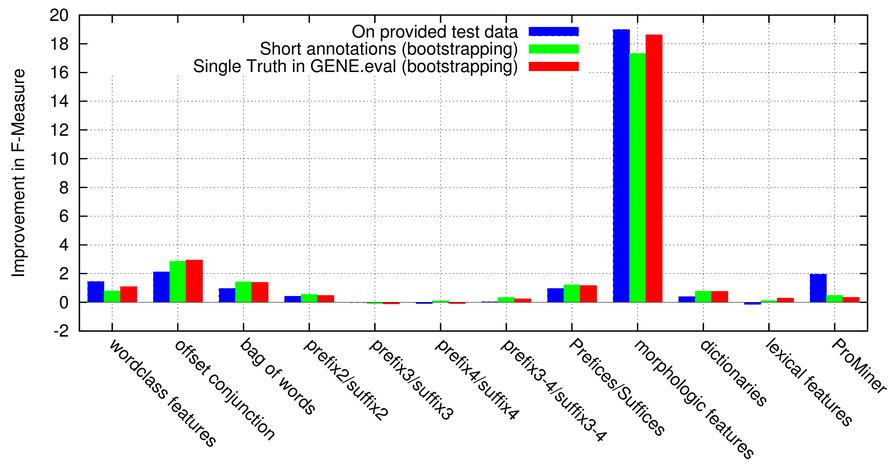
- Two Trainingsets:
  - Shortest possible annotation: Example (see 2nd sentence above): factor IX and factor IXa and factor XIa
  - Longest possible annotation: Example: human factor IX and factor IXa and factor XIa

How to deal with different taggings?

- Assume as example: ...fibrinogen degradation products (FDP)...
  - *Long* Annotator: fibrinogen degradation products
  - *Short* Annotator: fibrinogen ; FDP
- Use long annotation first, then add short annotation (without overlaps): fibrinogen degradation products and FDP
- Use short annotation first, then add long annotation (without overlaps): fibrinogen and FDP
- Greedy: Combine both (with overlaps): fibrinogen and FDP and fibrinogen degradation products

Model Selection

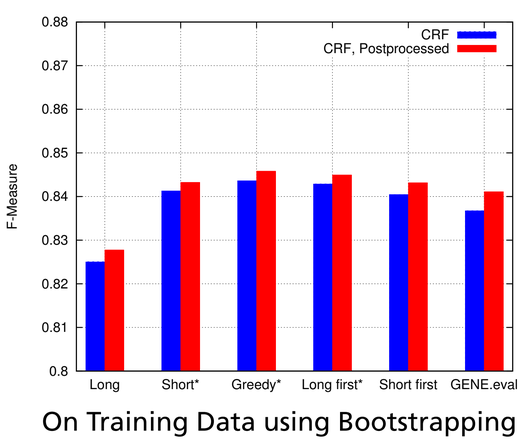
- Bootstrapping with 50 replicates
- Compared different tokenisations, impact is 2.48% on test data
- Rich set of features
  - Morphological, some automatically generated like bag-of-words, prefixes, suffixes, (brief) word class...
  - POS/Shallow Parsing: GeniaTagger
  - Annotations from ProMiner [1] as features
    → Very high precision because of mapping to UniProt and EntrezGene
  - Difficult to analyse optimal combination of features
    Example: prefixes with different length (see figure)



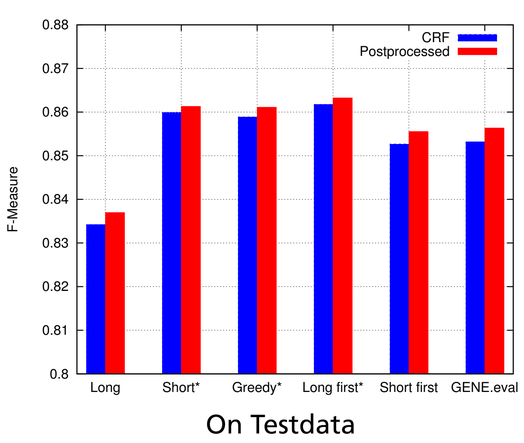
Results and Discussion

Model	Bootstrapping on Trainingset Precision		Recall		F-Score		On Testset Precision	Recall	F-Score
GENE.eval	86.61	(0.0071)	81.76	(0.0123)	84.11	(0.0076)	87.86	83.53	85.64
Long	86.30	(0.0065)	79.53	(0.0094)	82.78	(0.0064)	87.41	80.29	83.70
Short*	86.87	(0.0054)	81.94	(0.0106)	84.33	(0.0069)	88.57	83.83	86.13
Greedy*	80.21	(0.0069)	89.47	(0.0057)	84.58	(0.0047)	82.02	90.63	86.11
Long first*	85.38	(0.0060)	83.63	(0.0079)	84.50	(0.0055)	87.27	85.41	86.33
Short first	83.83	(0.0063)	84.81	(0.0065)	84.32	(0.0048)	85.50	85.61	85.56

(* submitted results)



On Training Data using Bootstrapping



On Testdata

- Short Annotation: best Precision
- Long Annotation: harder to find, but mostly matches author's mind

⇒ Good trade-off: Long first combination

- Greedy Combination: High Recall because of redundant annotation

⇒ good precondition for normalisation tasks

- Remarkable differences between results on test set using bootstrapping and training set are untypical

→ Impact of ProMiner?

References

[1] D. Hanisch, K. Fundel, H. T Mevissen., R. Zimmer, and J. Fluck. ProMiner: Organism-specific protein name detection using approximate string matching. *Proceedings of the BioCreative Challenge Evaluation Workshop 2004*, 2004.

[2] A. K. McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002.

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