Conditional Random Fields for Named Entity Recognition

Feature Selection and Optimization for Biology and Chemistry

Roman Klinger

http://www.roman-klinger.de/

30. Mai 2011

technische universität dortmund

Fraunhofer SCAI

Überblick

Motivation

Ausgangssituation:

- 1996:
  Namenserkennung von Personen-, Orts-, Organisationsnamen
  - → 97 % Präzision/96 % Vollständigkeit[1]
- 2005:
  Namenserkennung von Gen-/Proteinnamen
  - → 83 % Präzision/Vollständigkeit[2]
- Keine öffentliche Korpora
  für Chemienamen oder Mutationen etabliert

[1] Grishman, Sundheim: "Message Understanding Conference – 6: A Brief History. COLING 1996.

[2] Yeh et al.: "BioCreAtIvE Task IA: gene mention finding evaluation", BMC Bioinformatics 2005.

Motivation

Zielsetzung:

- Entwicklung vereinfachter Schritte zur Erstellung von Modellen zur Namenserkennung biologischer und chemischer Entitäten

Named Entity Recognition

- Gehört zur Informationsextraktion
- Problem:
  Markiere Terme in Freitext, welche zu einer vorgegebenen Klasse gehören
- Beispiele:
  Personennamen, chemische Namen, Gennamen, Einheiten
- Anwendungen:
  Information Retrieval, Relation Extraction, Semantic Search...

Named Entity Recognition – Beispiel

Markiere alle chemischen Namen, die der IUPAC[3]-Nomenklatur folgen

. . . on a large scale by reaction of the corresponding N-phenylhydrazones of 9-ethyl-3-carbazolecarbaldehyde, 9-ethyl-3,6-carbazoledicarbaldehyde, 4-dimethyl-amino-, 4-diethylamino-, 4-benzylethylamino-, 4-(diphenylamino)-, 4-(4,4-4′-dimethyl-diphenylamino)-, 4-(4-formyldiphenylamino)- and 4-(4-formyl-4′-methyldiphenyl-amino)benzaldehyde with. . .

[3] International Union of Pure and Applied Chemistry

Named Entity Recognition – Beispiel

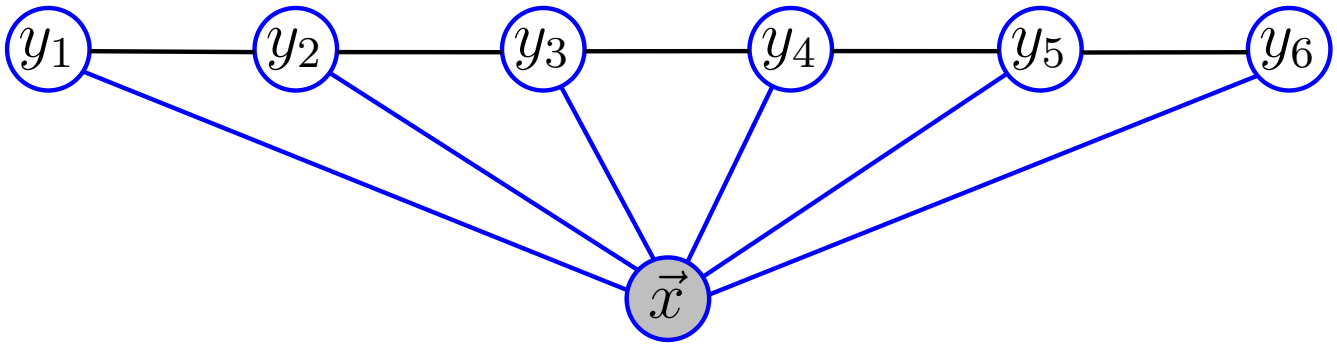
Markiere alle chemischen Namen, die der IUPAC[3]-Nomenklatur folgen

. . . on a large scale by reaction of the corresponding N-phenylhydrazones of 9-ethyl-3-carbazolecarbaldehyde, 9-ethyl-3,6-carbazoledicarbaldehyde, 4-dimethyl-amino-, 4-diethylamino-, 4-benzylethylamino-, 4-(diphenylamino)-, 4-(4,4-4′-dimethyl-diphenylamino)-, 4-(4-formyldiphenylamino)- and 4-(4-formyl-4′-methyldiphenyl-amino)benzaldehyde with. . .

[3] International Union of Pure and Applied Chemistry

Conditional Random Fields

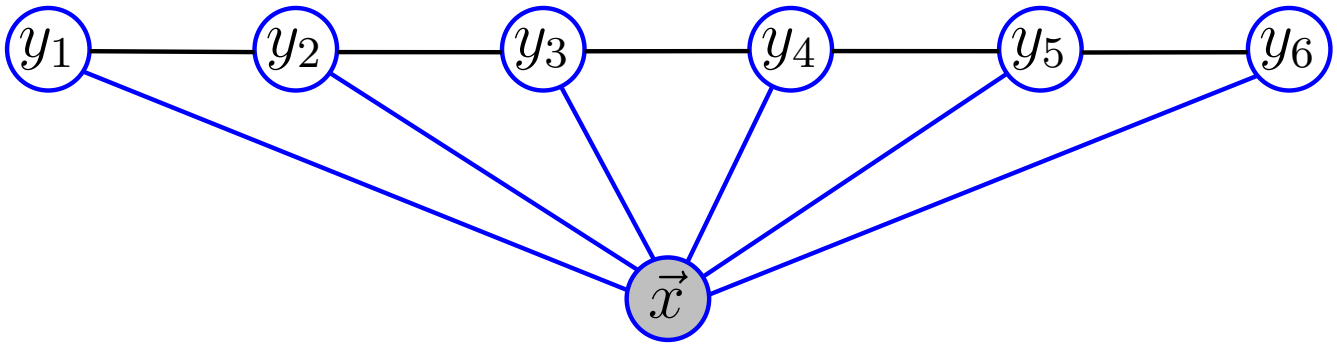
- Dekomposition einer bedingten Wahrscheinlichkeitsverteilung, typischerweise lineare Kette



- $$p_{\vec{\lambda}}(\vec{y}|\vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \cdot \exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m} \lambda_i f_i\left(y_{j-1}, y_j, \vec{x}, j\right)\right)$$

Conditional Random Fields

- Dekomposition einer bedingten Wahrscheinlichkeitsverteilung, typischerweise lineare Kette

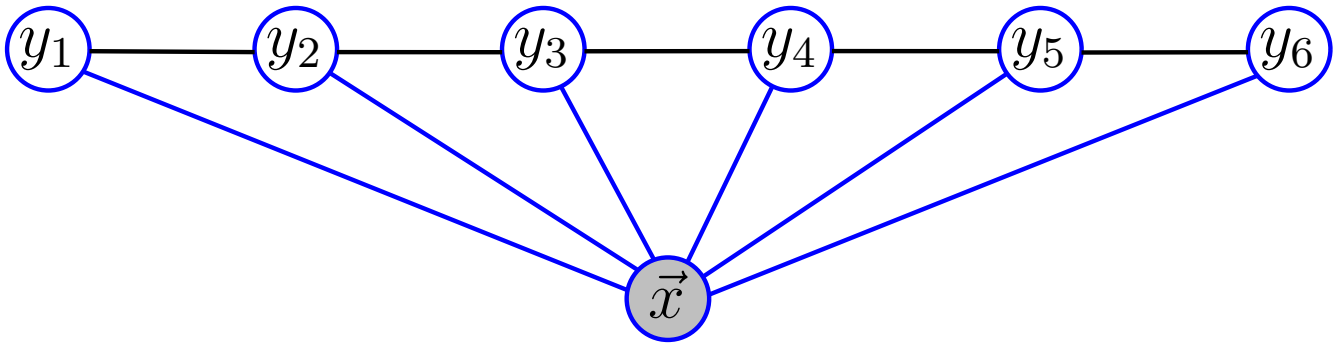


- $p_{\vec{\lambda}}(\vec{y}|\vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \cdot \exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m} \lambda_i f_i\left(y_{j-1}, y_j, \vec{x}, j\right)\right)$
- Anwendung zum Segmentieren von Text mittels IOB-Format

$\vec{x} =$	...	corresponding	N	-	phenylhydrazones	of	...
$\vec{y} =$	...	O	B	I	I	O	...

Conditional Random Fields

- Dekomposition einer bedingten Wahrscheinlichkeitsverteilung, typischerweise lineare Kette



- $p_{\vec{\lambda}}(\vec{y}|\vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \cdot \exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m} \lambda_i f_i\left(y_{j-1}, y_j, \vec{x}, j\right)\right)$
- Anwendung zum Segmentieren von Text mittels IOB-Format

$\vec{x} =$	...	corresponding	N	-	phenylhydrazones	of	...
$\vec{y} =$	...	O	B	I	I	O	...

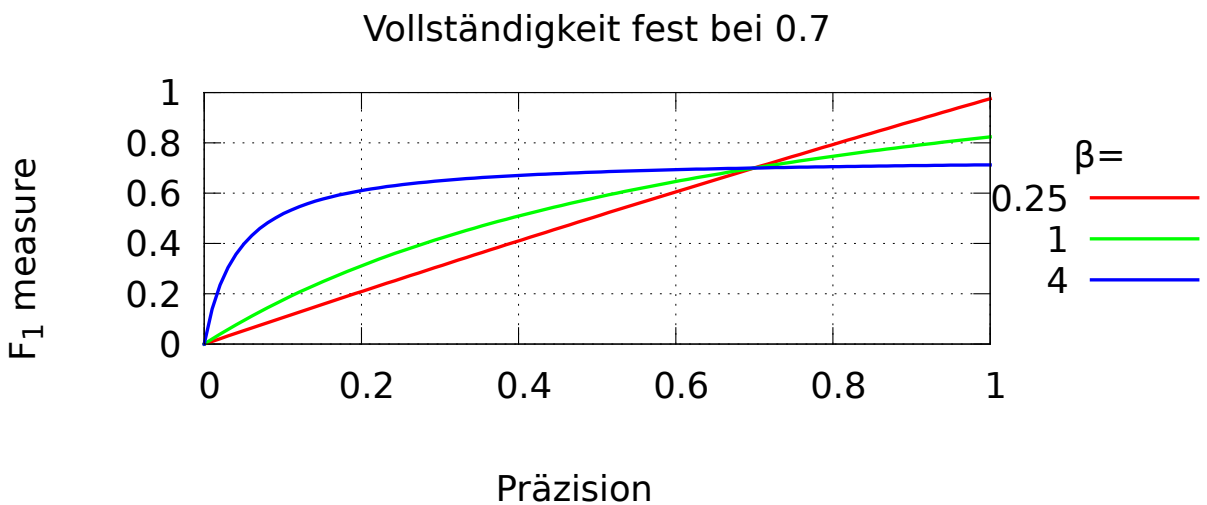
- Training: $\overset{!}{\max}_{\vec{\lambda}} \left(\log p_{\vec{\lambda}}(\vec{y}|\vec{x})\right)$, Inferenz: Viterbi

Evaluierung

- Üblicherweise Verwendung des F_β Maßes

$$F_\beta(\vec{\lambda}, \mathcal{D}) = \frac{(1+\beta^2) \cdot \mathsf{prec}(\vec{\lambda}, \mathcal{D}) \cdot \mathsf{rec}(\vec{\lambda}, \mathcal{D})}{\beta^2 \cdot \mathsf{prec}(\vec{\lambda}, \mathcal{D}) + \mathsf{rec}(\vec{\lambda}, \mathcal{D})} .$$

- Gewichtetes harmonisches Mittel von Vollständigkeit (recall) und Genauigkeit (precision), wobei
  - $\mathsf{prec}(\vec{\lambda}, \mathcal{D}) = \frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FP}}$ und $\mathsf{rec}(\vec{\lambda}, \mathcal{D}) = \frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}}$



Überblick über die Dissertation

Überblick über die Dissertation

Erstellen der notwendigen Schritte zur Erkennung von Entitätsklassen aus den Lebenswissenschaften:

- Gen- und Proteinnamen
- Nennungen von Mutationen
- Chemischen Namen, die der IUPAC-Nomenklatur folgen

Hierbei jeweils:

- Diskussion und Erstellung von Trainingskorpora
- Diskussion spezifischer Probleme
- Modellselektion

Überblick über die Dissertation

Hierbei wurde des Weiteren betrachtet:

- Einbinden der Information von mehreren Annotatoren
- Temporale Stabilität eines Modells
- Normalisierung der gefundenen Entitäten
- Vergleich mit anderen Verfahren

Überblick über die Dissertation

Erweiterungen um die Erstellung von Modellen zu vereinfachen und die Anwendbarkeit zu verbessern:

- Mehrkriterielle Optimierung zum Training von CRFs
  - Wahl von Präzision- und Vollständigkeit-Gewichtung ohne Verschlechterung der Laufzeit

Überblick über die Dissertation

Erweiterungen um die Erstellung von Modellen zu vereinfachen und die Anwendbarkeit zu verbessern:

- Mehrkriterielle Optimierung zum Training von CRFs
  - Wahl von Präzision- und Vollständigkeit-Gewichtung ohne Verschlechterung der Laufzeit
- Automatisierte Suche nach performanzsteigernden Strukturen (im Vergleich zu linearen Ketten)
  - Betrachtung von entfernten Relationen zwischen Variablen

Überblick über die Dissertation

Erweiterungen um die Erstellung von Modellen zu vereinfachen und die Anwendbarkeit zu verbessern:

- Mehrkriterielle Optimierung zum Training von CRFs
  - Wahl von Präzision- und Vollständigkeit-Gewichtung ohne Verschlechterung der Laufzeit
- Automatisierte Suche nach performanzsteigernden Strukturen (im Vergleich zu linearen Ketten)
  - Betrachtung von entfernten Relationen zwischen Variablen
- Merkmalsselektion zur Steigerung von Generalisierbarkeit und Geschwindigkeit
  - Adaption von Verfahren aus Klassifikationsproblemen

Erkennung von IUPAC-Namen

Erkennung von IUPAC-Namen

Besondere Herausforderung:

- Unendlich viele Namen denkbar, da nach Regeln erstellt
- Regeln werden nur grob befolgt und mit anderen Nomenklaturen gemischt
- Namen erschweren die Erkennung der Grammatik
- Erkennung von Beginn und Ende eines Begriffs kritisch
- Verarbeitung solcher Begriffe beim Verleger erzeugt Fehler
- Tokenisierung fraglich

Weiteres Beispiel:

18-bromo-12-butyl-11-chloro-4,8-diethyl-5-hydroxy-15-methoxytricos-6,13-dien-19-yne-3,9-dione

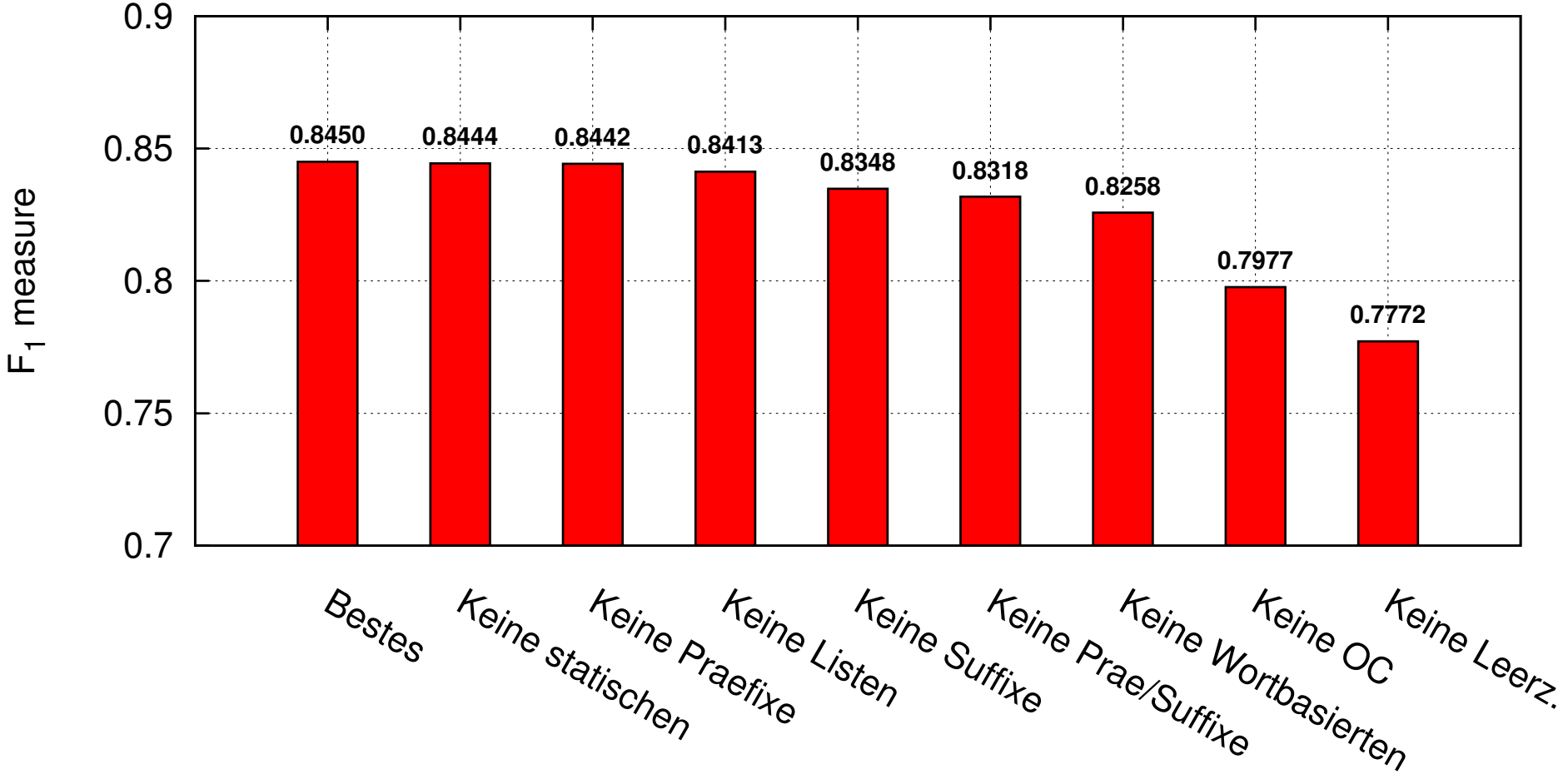
Merkmale: Wie werden Token repräsentiert?

- Statische Merkmale, z.B.
  - Großbuchstaben (AMPA)
  - Bindestriche (- – — ―)
  - Klammern (()[]{})
- Automatisch erstellte, z.B.
  - Präfix, Suffix, Wort (pyrimidine)
- Kontextuelle, z.B.
  - Weißraum vor/nach Token (_pyrimidine_)
  - Kombination der Merkmale sukzessiver Tokens
- Wörterbuch-basierte, z.B.
  - Präfix/Suffix-Listen

Erzeugt über 500.000 Merkmale!

Modellselektion (via Bootstrapping), Ergebnis

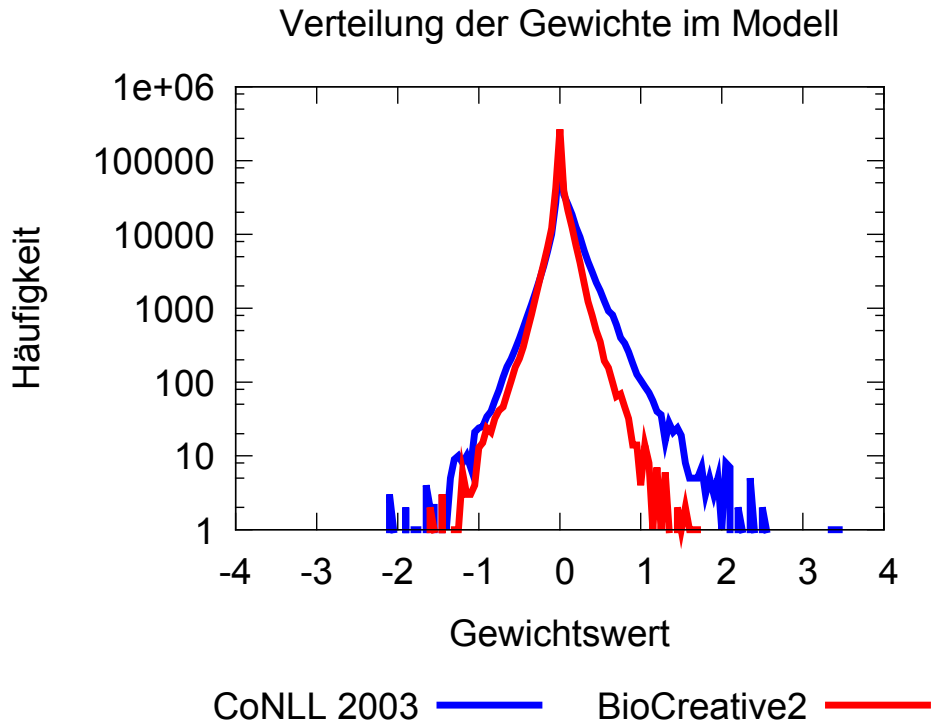
Welche Merkmale sind wirklich wichtig?



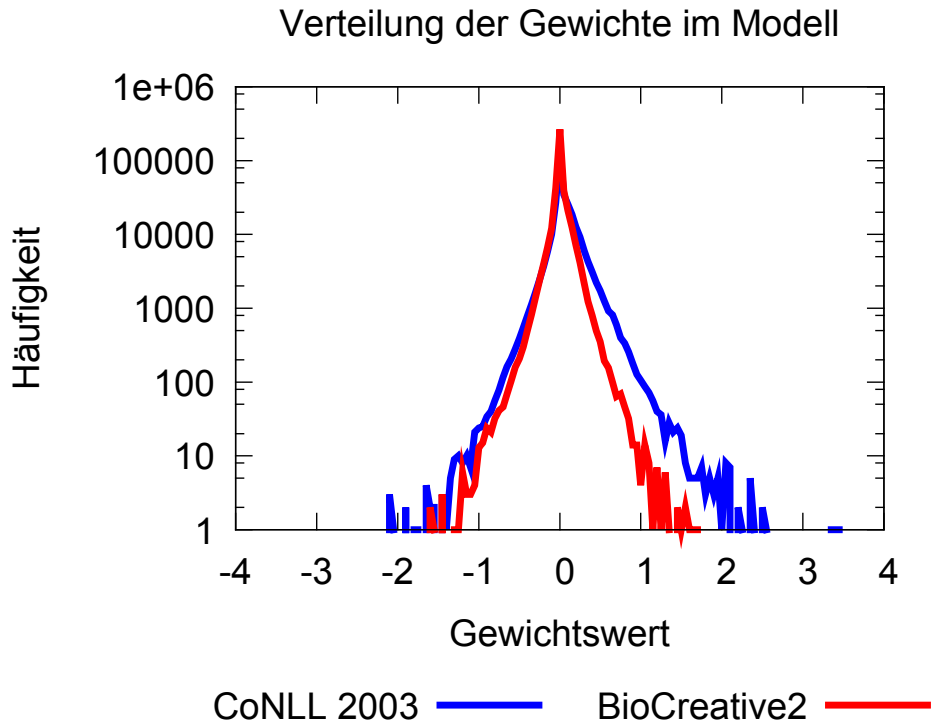
Test Korpus: 85,6 % F_1 (86,5 % P., 84,8 % R.)

Merkmalsselektion

Merkmalsselektion: Motivation



Merkmalsselektion: Motivation



- Viele Merkmale haben wahrscheinlich wenig Einfluß
- Eingeschränkte Generalisierbarkeit
- Zusätzlicher Aufwand mit
  - Merkmalsextraktion
  - Speicherbedarf
  - Zeit

Merkmalsselektion: Motivation

Laufzeit einer Iteration des CRF Trainings oder Inferenz:

$$\mathcal{O}(l^2 nm)$$

- l Anzahl verschiedener Labels
- n Anzahl von Faktoren
- m Durchschnittliche Anzahl von Merkmalen je Faktor

- Spärliche Nutzung von Merkmalen (auf Testmenge)
  CoNLL $m = 7,462$, IUPAC $m = 0,53$, BC2 $m = 7,47$
- Verwaltung der Datenstrukturen zusätzlich aufwendig

Vorherige Arbeiten

- Iterative Konstruktion von Merkmalsverknüpfungen

A. McCallum.
Efficiently Inducing Features of Conditional Random Fields.
In *Proc. of the 19th Conference in Uncertainty in Articifical Intelligence (UAI-2003)*, pages 403–410, 2003.

- Analyse von Straftermen zur Regularisierung

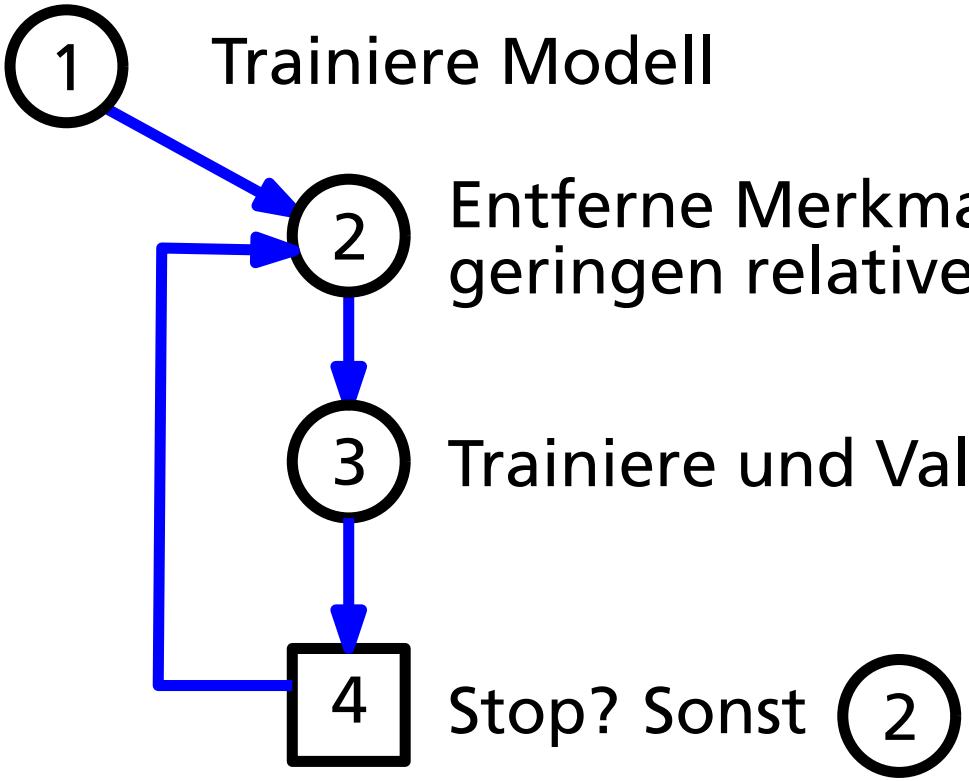
F. Peng and A. McCallum.
Accurate Information Extraction from Research Papers using Conditional Random Fields.
In *Proc. of HLT-NAACL*, pages 329–336, 2004

- L_1 norm Regularisierung

D. L. Vail, J. D. Lafferty, and M. M. Veloso.
Feature Selection in Conditional Random Fields for Activity Recognition.
In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3379–3384, 2007.

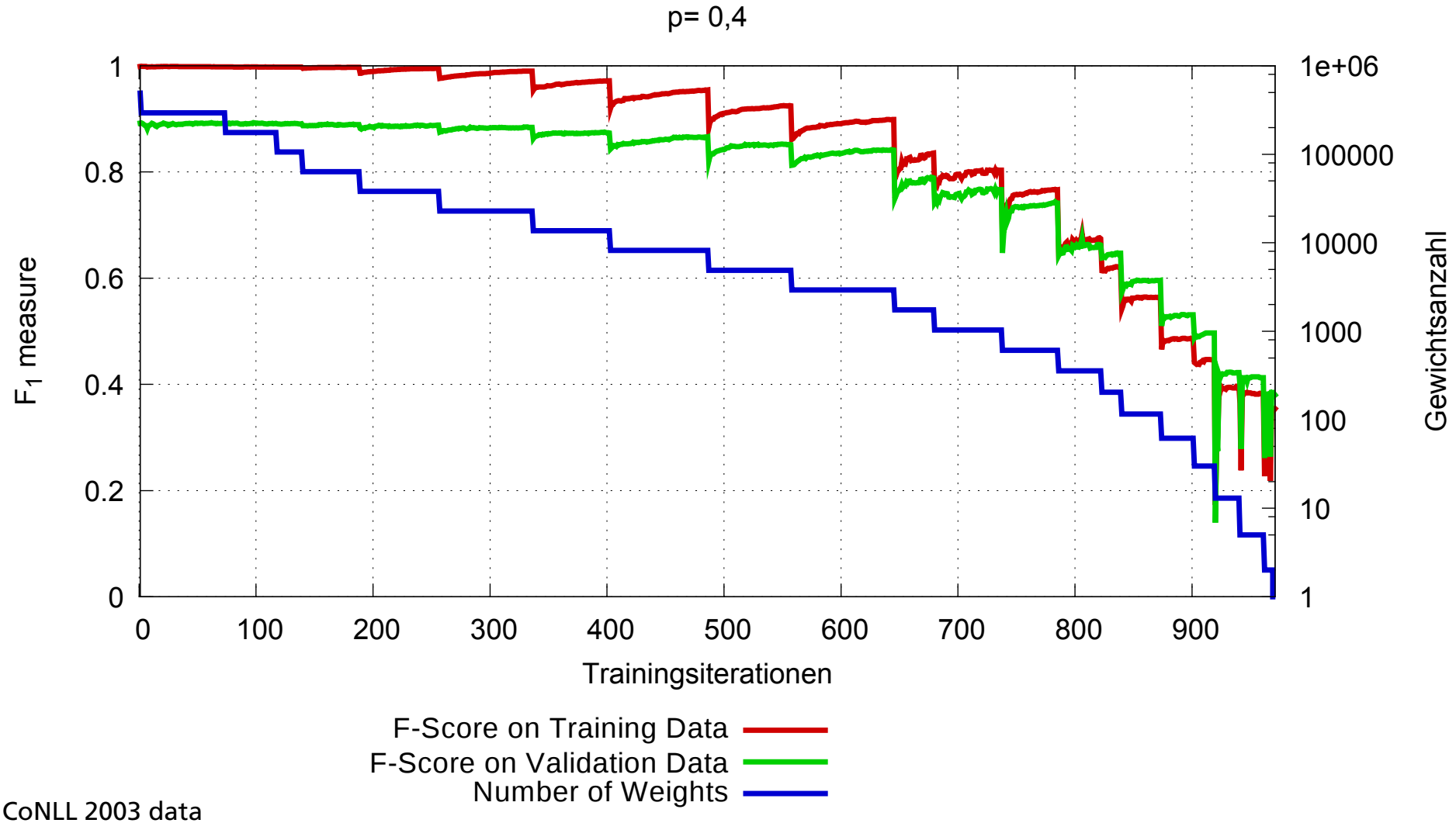
$\Longrightarrow$ Alle Methoden nutzen Trainingsverfahren
Hier: Filteransätze sowie ein iteratives Verfahren

Iterative Feature Pruning



- Selektiere beste Merkmalsmenge durch Kreuzvalidierung
- Parameter: Anteil p gelöschter Merkmale in jeder Iteration

Iterative Feature Pruning



CoNLL 2003 data

Iterative Feature Pruning – Zusammenfassung

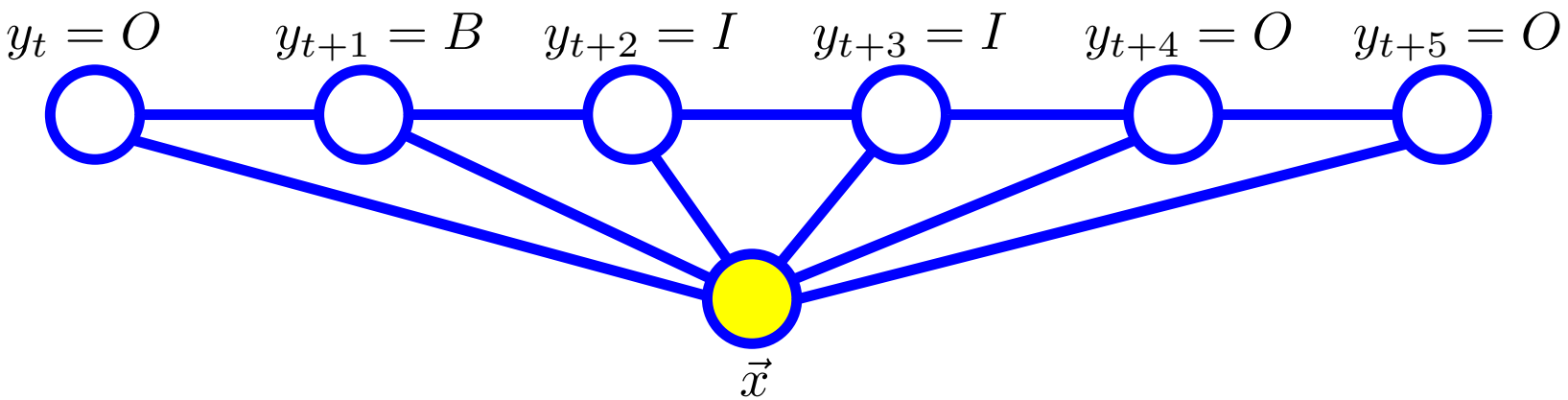
\+ Selektiert wichtige Merkmale unabhängig vom Übergang

\+ Niedrige Anzahl von Merkmalen: bedeutsam und verständlich

− Langsam, da Trainingsverfahren eingebunden ist

Filter

Teile Sequenz in Klassifikationsinstanzen

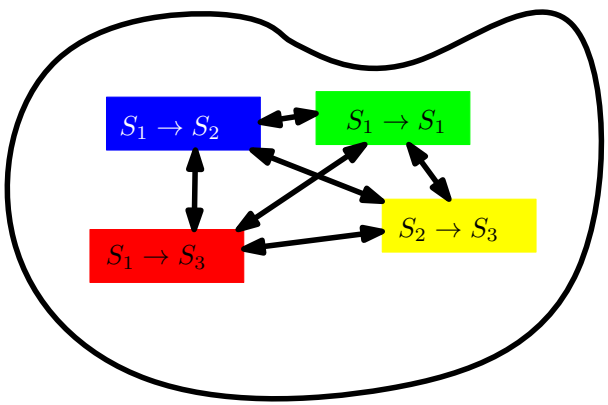


- Generiere eine Instanz aus jeder Clique
- Klassen sind:
  $O \to B$, $B \to I$, $I \to I$, $I \to O$, $O \to O$, ...

Filter Ansätze

Selektiere die wichtigsten Merkmale anhand von:

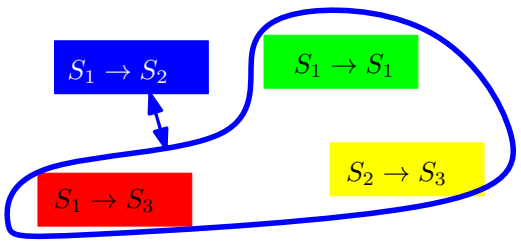
- Informationsgewinn
  - Bedeutung der Merkmale betreffend alle Transitionen



Filter Ansätze

Selektiere die wichtigsten Merkmale anhand von:

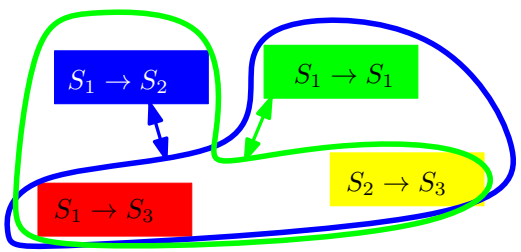
- Informationsgewinn
  - Bedeutung der Merkmale betreffend alle Transitionen
- Informationsgewinn (Einer-Gegen-Alle)
  - Bedeutung der Merkmale einer Transition gegen alle anderen Transitionen



Filter Ansätze

Selektiere die wichtigsten Merkmale anhand von:

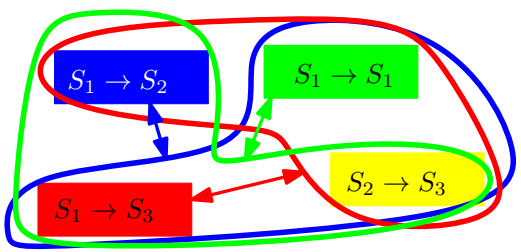
- Informationsgewinn
  - Bedeutung der Merkmale betreffend alle Transitionen
- Informationsgewinn (Einer-Gegen-Alle)
  - Bedeutung der Merkmale einer Transition gegen alle anderen Transitionen



Filter Ansätze

Selektiere die wichtigsten Merkmale anhand von:

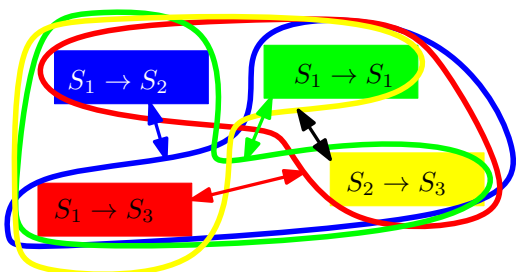
- Informationsgewinn
  - Bedeutung der Merkmale betreffend alle Transitionen
- Informationsgewinn (Einer-Gegen-Alle)
  - Bedeutung der Merkmale einer Transition gegen alle anderen Transitionen



Filter Ansätze

Selektiere die wichtigsten Merkmale anhand von:

- Informationsgewinn
  - Bedeutung der Merkmale betreffend alle Transitionen
- Informationsgewinn (Einer-Gegen-Alle)
  - Bedeutung der Merkmale einer Transition gegen alle anderen Transitionen



Filter Ansätze

Selektiere die wichtigsten Merkmale anhand von:

- Informationsgewinn
  - Bedeutung der Merkmale betreffend alle Transitionen
- Informationsgewinn (Einer-Gegen-Alle)
  - Bedeutung der Merkmale einer Transition gegen alle anderen Transitionen
- χ^2 (Einer-Gegen-Alle)
  - Test der Merkmale einer Transition gegen alle anderen Transitionen

Filter Ansätze

Selektiere die wichtigsten Merkmale anhand von:

- Informationsgewinn
  - Bedeutung der Merkmale betreffend alle Transitionen
- Informationsgewinn (Einer-Gegen-Alle)
  - Bedeutung der Merkmale einer Transition gegen alle anderen Transitionen
- χ^2 (Einer-Gegen-Alle)
  - Test der Merkmale einer Transition gegen alle anderen Transitionen
- Zufall zum Vergleich als untere Schranke

Filter Ansätze

Selektiere die wichtigsten Merkmale anhand von:

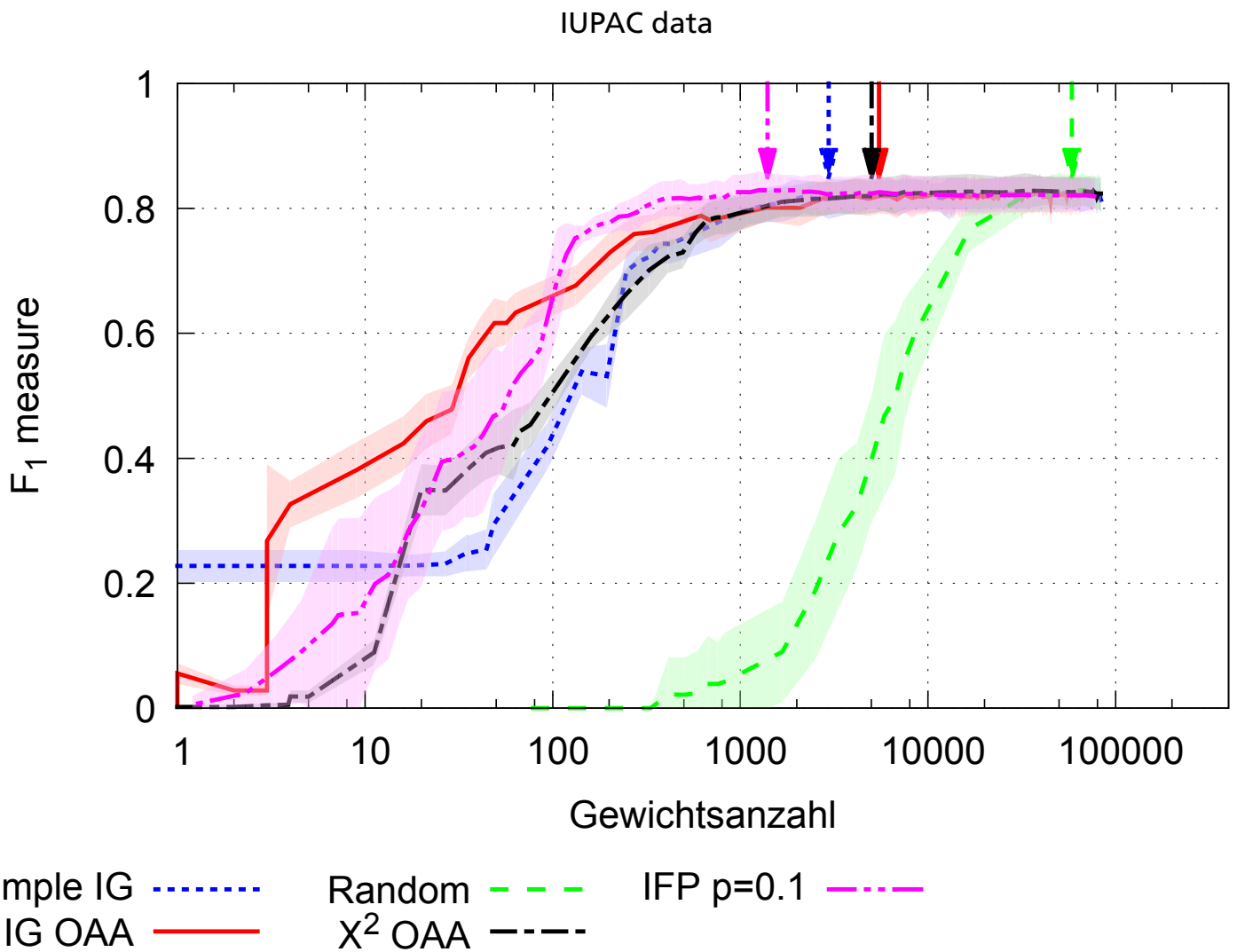
- Informationsgewinn
  - Bedeutung der Merkmale betreffend alle Transitionen
- Informationsgewinn (Einer-Gegen-Alle)
  - Bedeutung der Merkmale einer Transition gegen alle anderen Transitionen
- χ^2 (Einer-Gegen-Alle)
  - Test der Merkmale einer Transition gegen alle anderen Transitionen
- Zufall zum Vergleich als untere Schranke

Parameter: Anteil verbleibender Merkmale

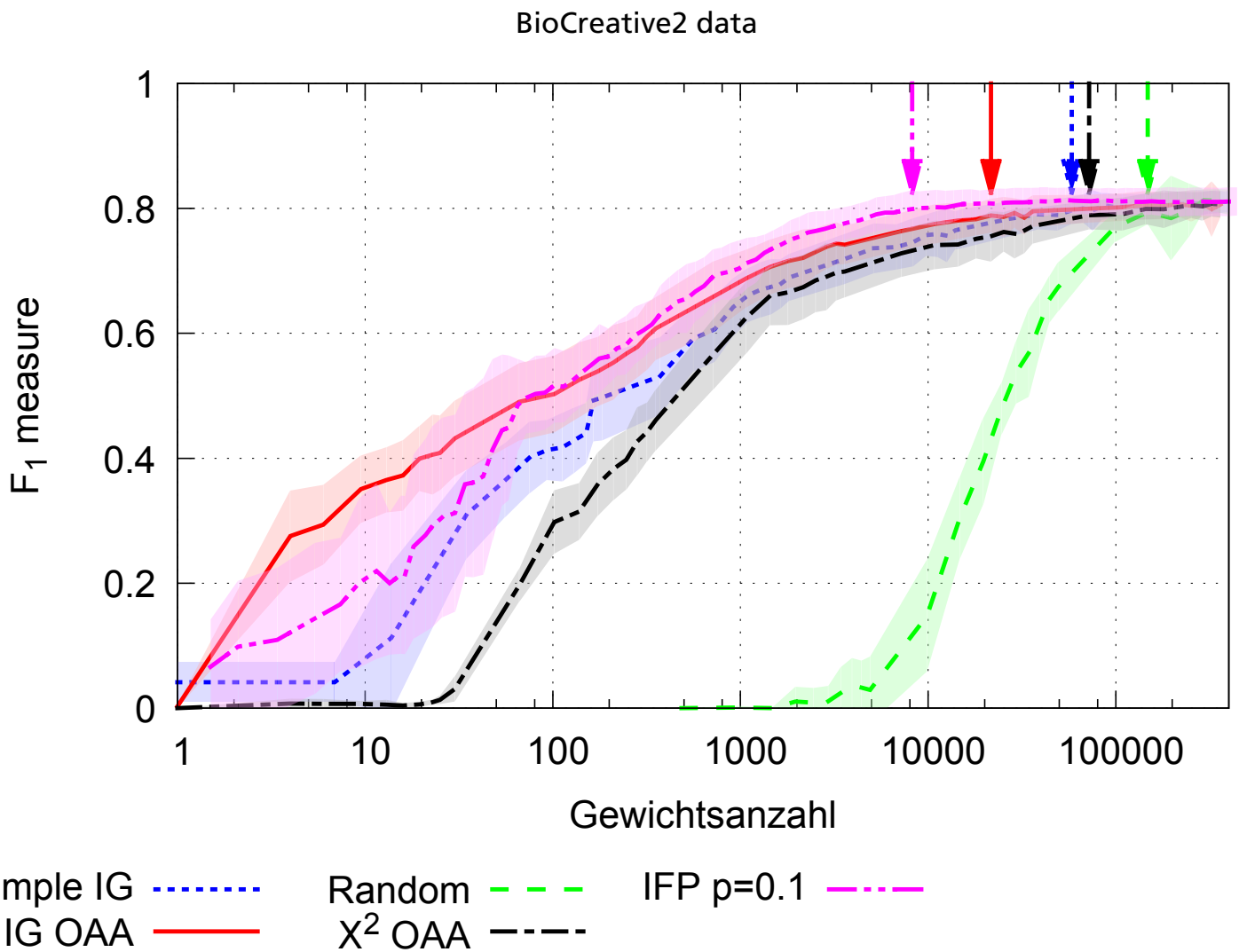
Filter – Zusammenfassung

+ Selektiert die Merkmale sehr schnell
+ Unabhängig vom Training
− Extrahierte Merkmale können korrelieren
− Alle Transitionen behalten dieselbe Anzahl von Merkmalen

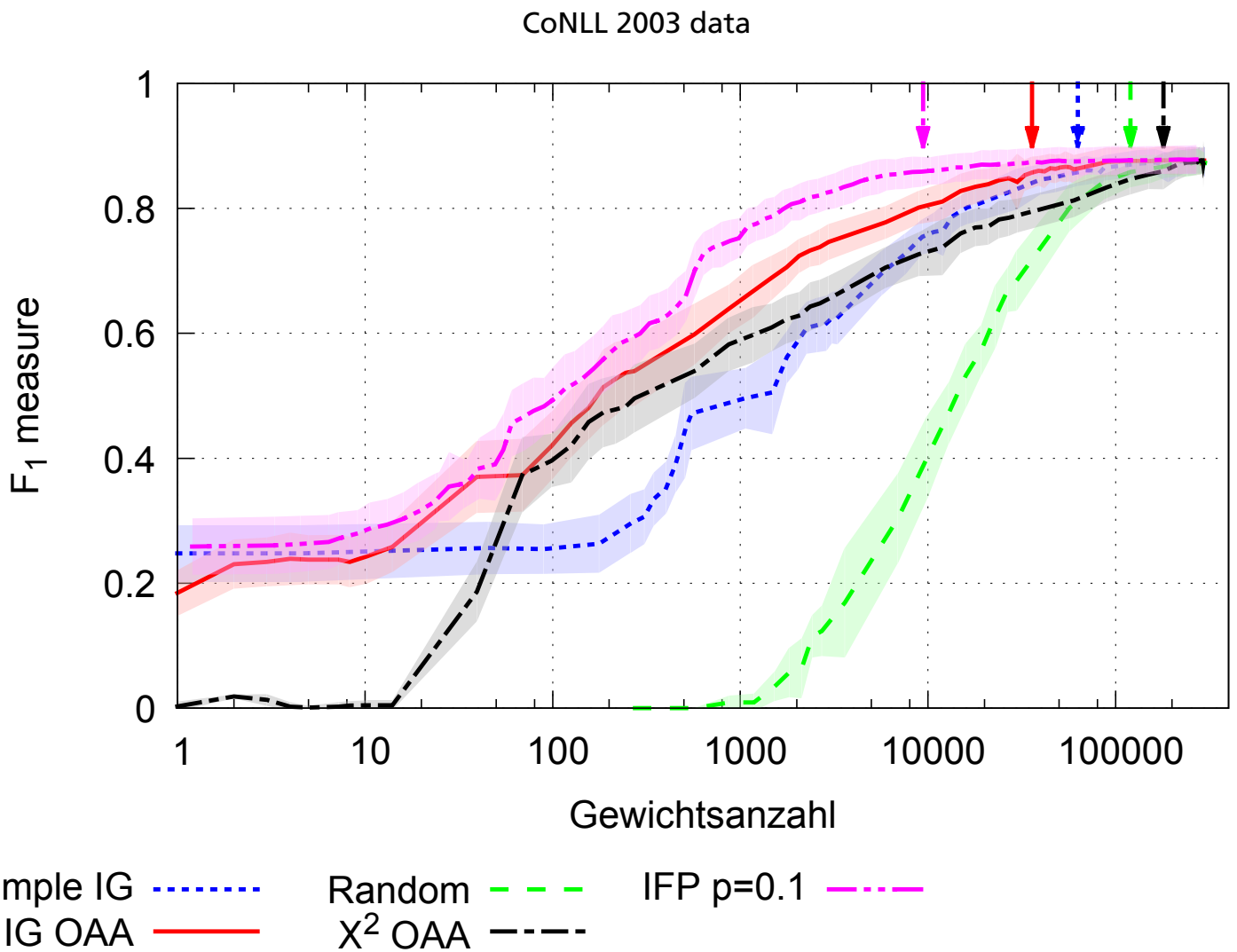
Parameterselektion via Kreuzvalidierung



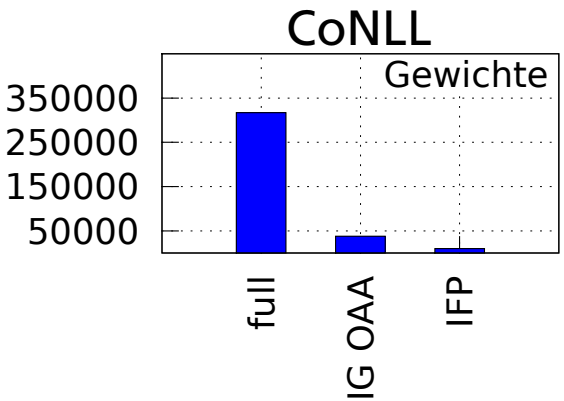
Parameterselektion via Kreuzvalidierung

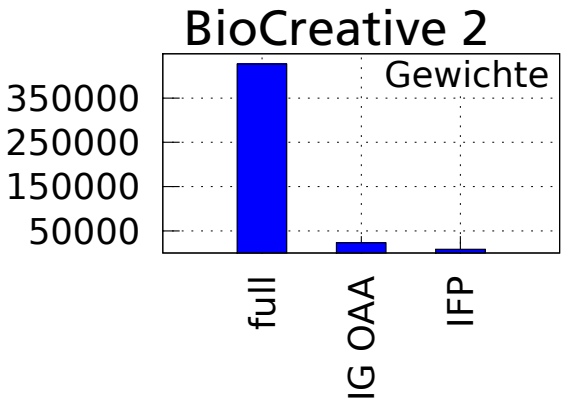


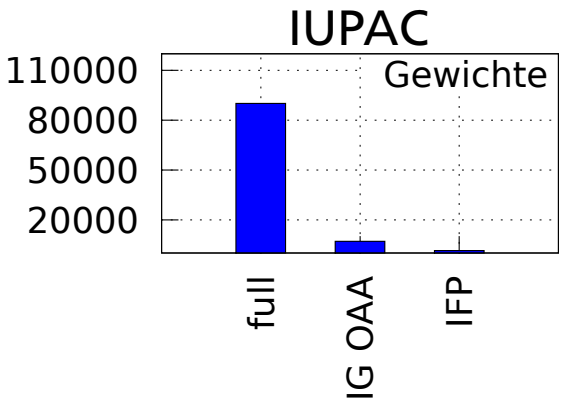
Parameterselektion via Kreuzvalidierung



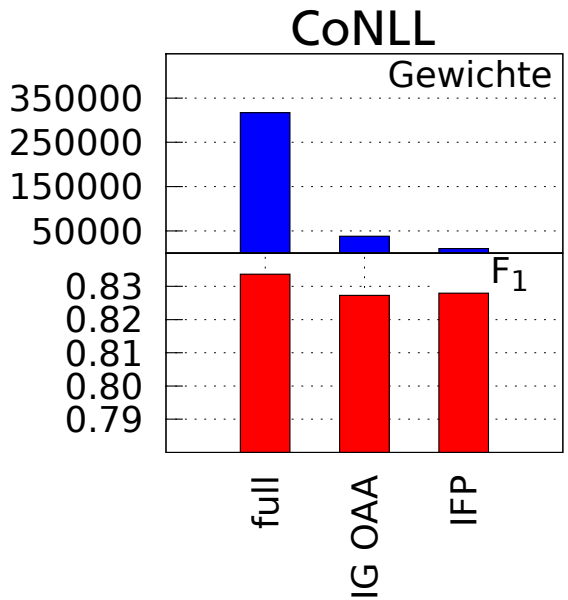
Test auf unabhängigen Mengen

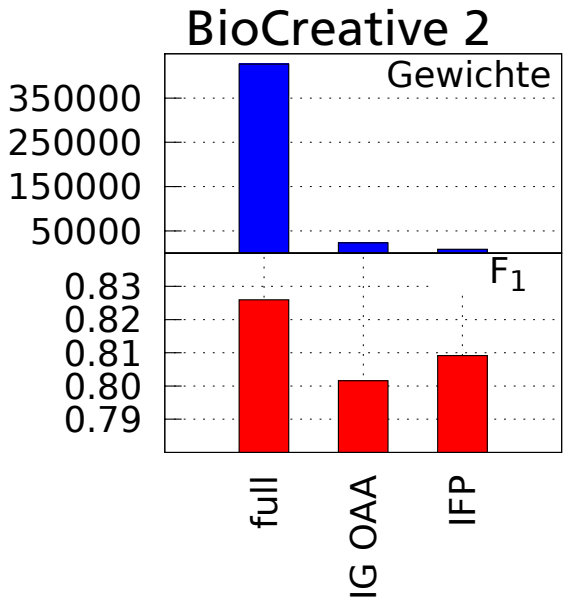


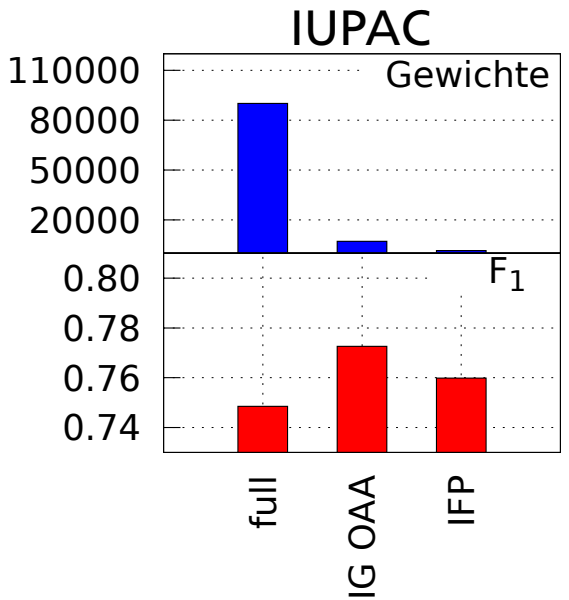




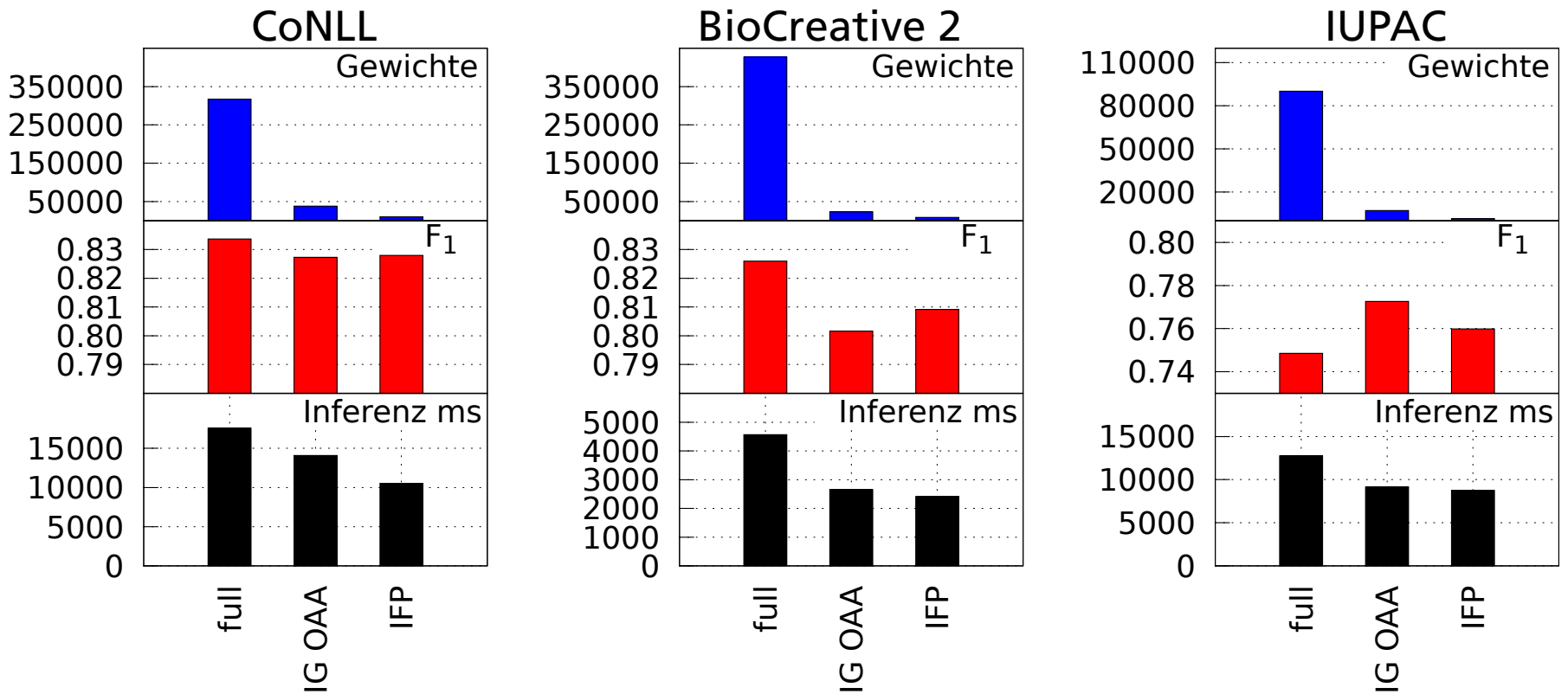
Test auf unabhängigen Mengen



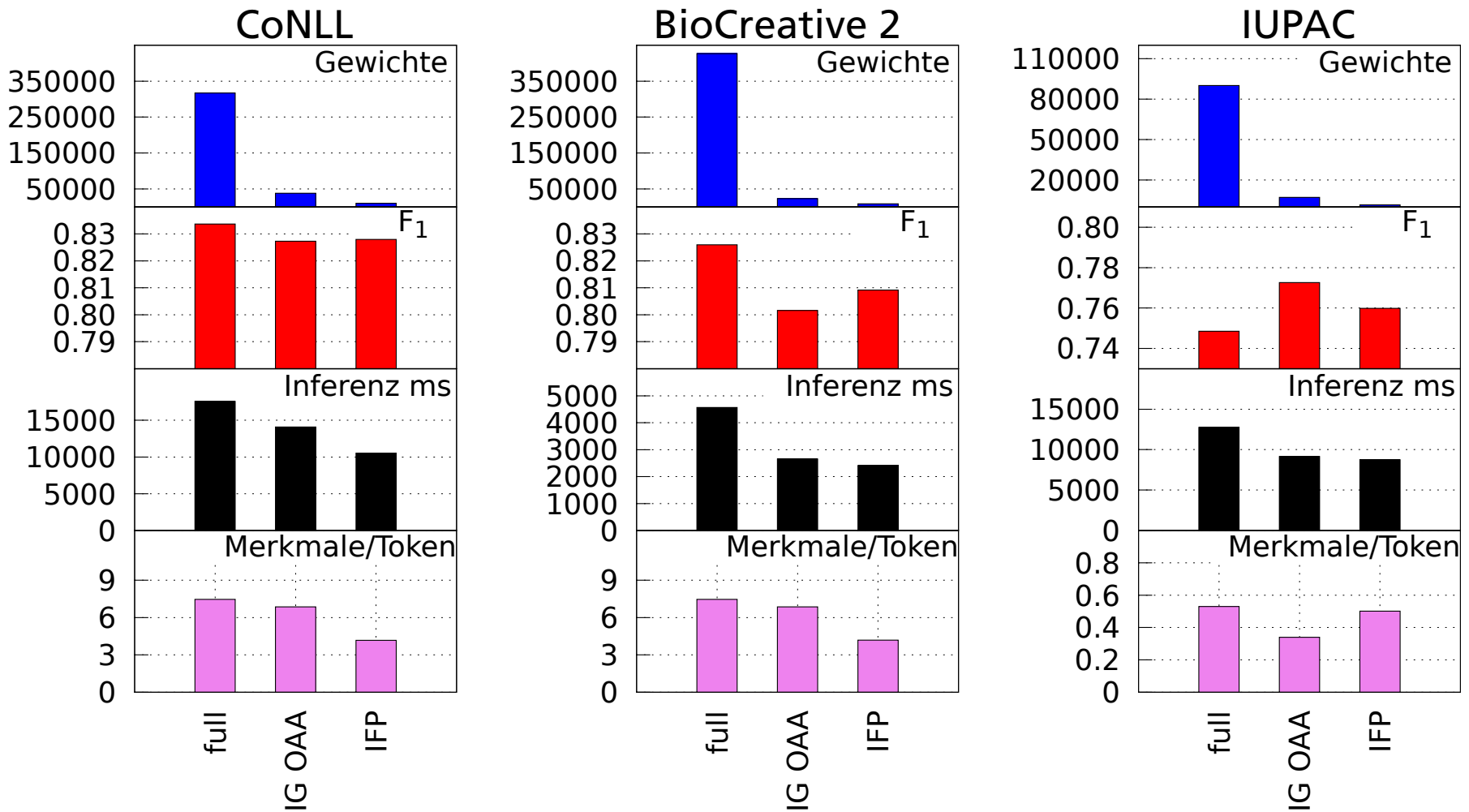




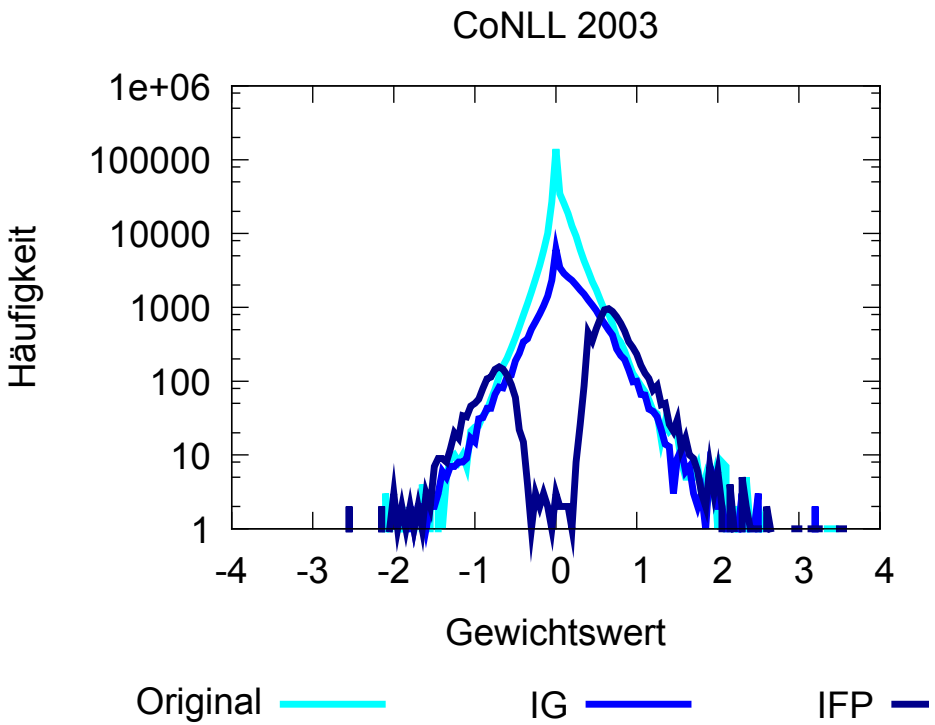
Test auf unabhängigen Mengen

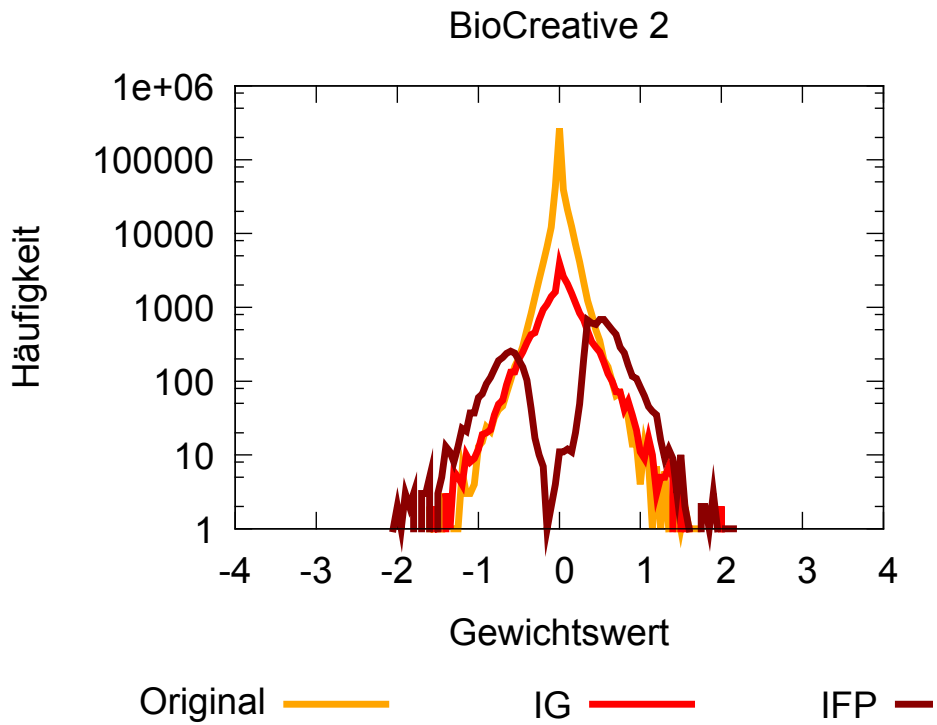


Test auf unabhängigen Mengen



Verteilung der Gewichte





Zusammenfassung

Zusammenfassung

- Adaptierte Merkmalsselektion zeigt deutlichen Geschwindigkeitsvorteil um 50 %
- Reduktion der Merkmale $\rightarrow$ 1,6 % (IUPAC) bis 3 % (Gene)
- Ermöglichung anderer Trainingsverfahren

Zusammenfassung

- System zur Erkennung von IUPAC-Namen erreicht 86 % F_1
- Gemeinsam mit den anderen Verfahren sind Schritte erarbeitet worden, die die Erstellung von Modellen für neue Entitätsklassen vereinfachen und die Anwendungsmöglichkeiten erweitern.