Gene Name Normalization at BioCreative Challenge 2

09.06.2008

Hauptseminar
Information Extraction in the Biomedical Domain
Summer Semester 2008
PD Dr. rer. nat. Günter Neumann

Speaker:
Stefan Fischer

Overview

- Motivation & Introduction
- BioCreAtIvE II Challenge
- Participants
  - ProMiner (RB)
  - Massively RB system
  - BioTagger (ML)
  - Me and my friends (semantic information)
- Conclusion

Motivation

- Huge amount of biomedical literature that cannot be handled manually.
- IE systems try to make this data accessible to biological experts and bioinformatics methods.
  - Literature network graphs
  - Summary of genes discussed in a text
    - Named entity recognition is not enough

Problems with NER

- Nomenclature
  - Evolved over time
  - Authors deviate from a recommended nomenclature
  - Or no standard at all
- Effects on gene names
  - Several synonymous aliases for one gene
  - Functionally unrelated genes share the same name
  - Permutations in multi-word names
  - Case-sensitive names
  - Overlap between gene names and general English words

Gene Normalization

- Tries to solve this problems by finding unique identifiers for mentions of gene names in a text.
- There are several approaches, but they are not comparable, because the creation of test sets is expensive.

2nd BioCreAtIvE (2006)

- ... Critical Assessment for Information Extraction in Biology
- Aim is to provide a framework for the construction of 'gold standard' data sets to train and test IE systems in biology.

- Tasks:
  - Gene mention tagging (last presentation)
  - Gene normalization
  - Extraction of protein-protein interactions from text

Gene Normalization Task

- Identify unique Entrez Gene identifiers for mentions of human genes and proteins in a MEDLINE abstract.
- Create a list of Entrez Gene IDs for each abstract in the test set.
- Simplifications:
  - Abstracts rather than full articles
  - Organism specific (human)
  - All mentions will be identified (relevant or not)

Data Preparation

- PubMed articles likely to have mentions of human genes and proteins. (Gene Ontology)
- 2 manual annotators, ~90% agreement
  - Training set (281 fully annotated abstracts)
  - Test set (262 fully annotated abstracts)
- Gene Ontology Annotation
  - Noisy training set (5,000 sparsely annotated abstracts, only relevant mentions)

Lexicon

- Entrez Gene identifier
- Names and aliases from NCBI, UniProt, HGNC
- Expansion with suffixes containing
  - „_HUMAN“, „1_HUMAN“ „H_HUMAN“
  - „protein“ „precursor“ „antigen“
- Removal of 381 most frequent terms
  - Unlikely to be gene names
  - „recessive“, „neural“, „liver“, „glycine“, „mediator“

⇒ 32,975 EntrezGene IDs with 163,478 synonyms

Scoring

- Simple matching of submitted list against gold standard
  - Submitted ID in gold standard → TP
  - Submitted ID not in gold standard → FP
  - Gold standard ID not in submitted list→ FN

- Ranking of teams by F-measure
  - Recall = TP/(TP+FN)
  - Precision = TP/(TP+FP)
  - F-measure = 2*P*R/(P+R)

ProMiner

Fraunhofer Institute

LMU München

ProMiner

- Search tool for gene and protein names in scientific publications
- Generation of disease centric databases
  - Auto Immune Data Base, @neurIST

- Rule-based
- Large curated, regularly updated dictionaries
- Token-based search algorithm
- Parenthesis expressions

ProMiner

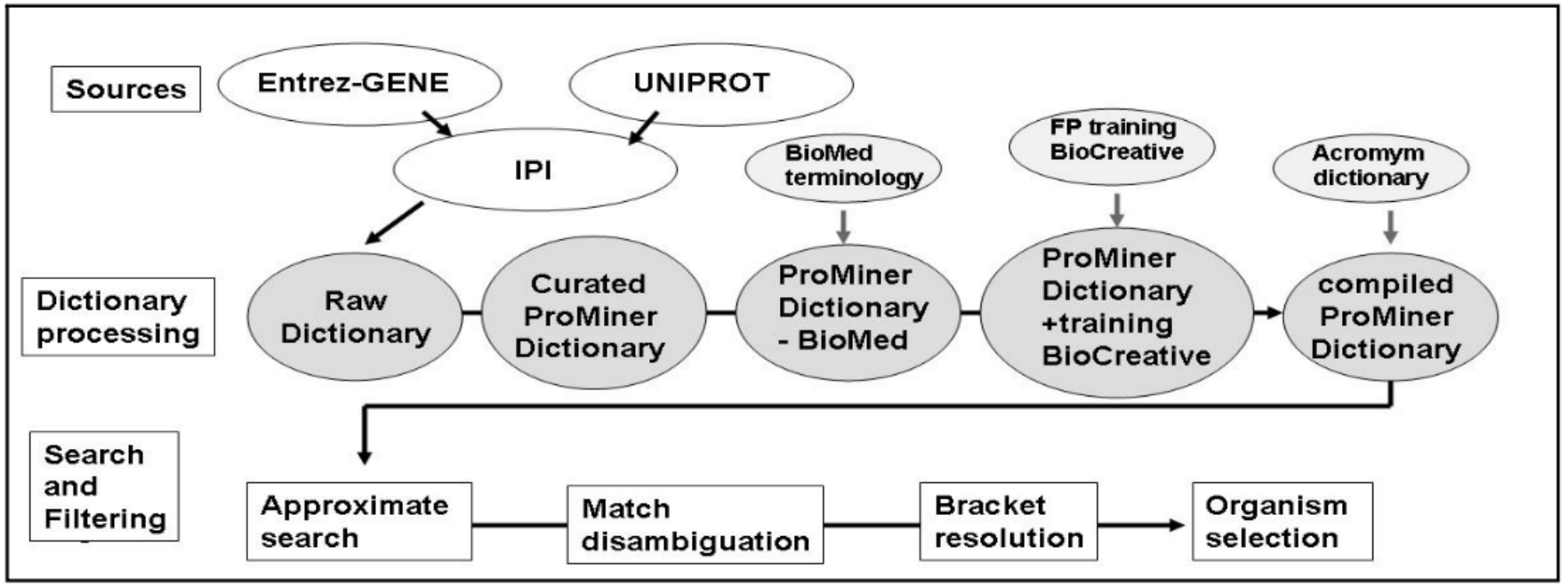


Figure 1: The ProMiner system used in BioCreAtIvE gene normalization

Dictionary sources

- EntrezGene
  - Gene description fields of human entries
- UniProt
  - Protein description fields of human entries
- IPI (International Protein Index)
  - Entries that are transitively mapped on IPI are merged into one dictionary entry

ProMiner

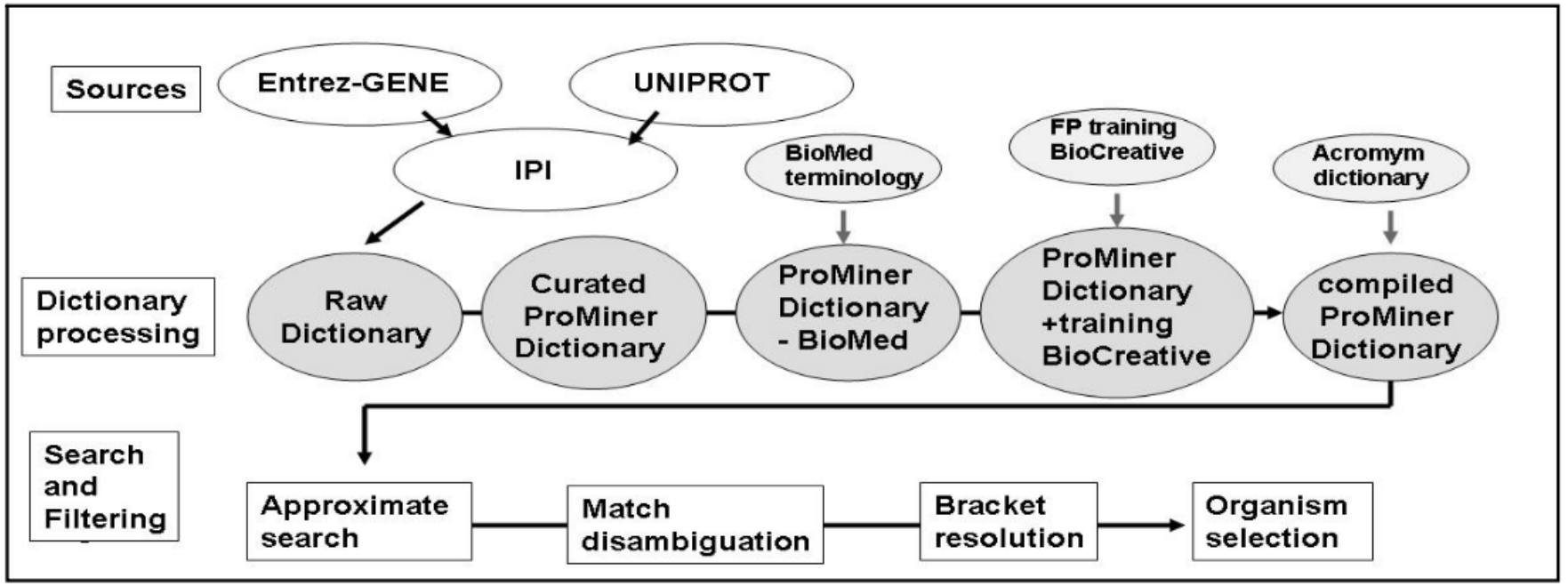


Figure 1: The ProMiner system used in BioCreAtIvE gene normalization

Automatic dictionary curation

- Acronym expansion (IL → Interleukin)
  - adding long-forms to dictionary
- Adding of spelling variants („IL 1“ → „IL1“)
- One-word synonyms
  - leading „h“ (SMRP → hSMRP, only if unique)
- Subtype specifiers (a → alpha)

⇒ Higher recall

- Filtering of unspecific synonyms with RE
  - d* M → „35 kDa protein“
- Manually curated list from other projects (Auto Immune Data Base)
  - Family names (‚membrane protein‘)
  - Physical descriptions („cDNA clone“, „5'end“)

ProMiner

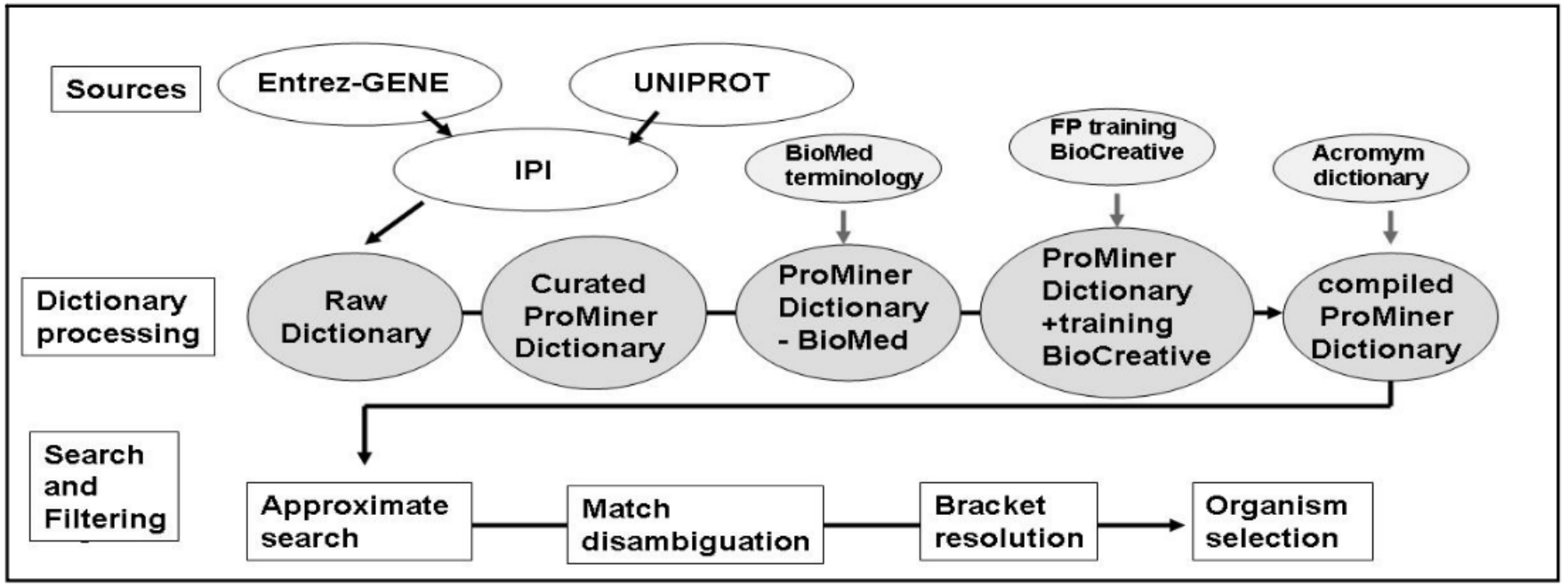


Figure 1: The ProMiner system used in BioCreAtIvE gene normalization

Curation & Training

- Removal of unspecific BioMed terminology
  - Open Biomedical Ontology
  - disease, tissue, organism and protein family names
- Training for BioCreAtIvE II
  - False Positives from training and noisy data
  - Inspection by an expert → curation list

Acronym dictionary

- Acronyms in the dictionary
  - Biomedical Abbreviation Server
- Pattern matching on all MEDLINE abstracts
  - „… respiratory distress syndrome (RDS) …“
- Reduction to acronyms similar to gene names
- Removal of long forms = dictionary entry

⇒ Gene search specific acronym dictionary

ProMiner

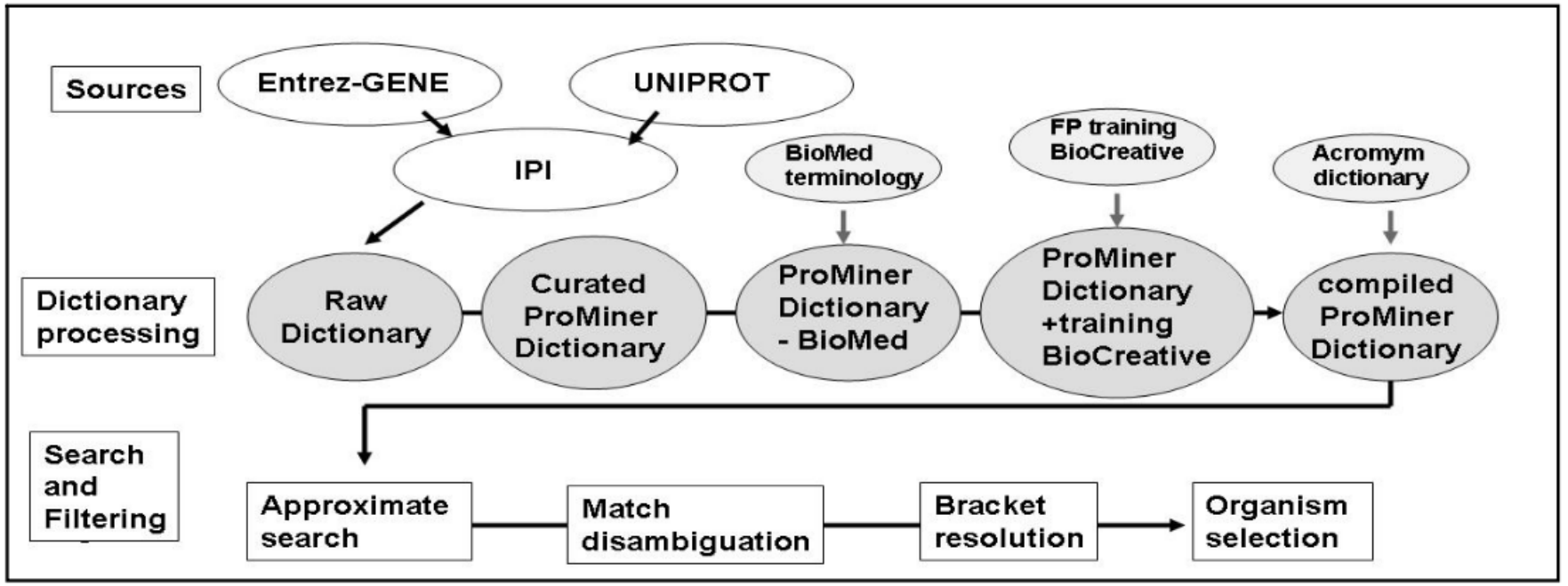


Figure 1: The ProMiner system used in BioCreAtIvE gene normalization

Compilation step

- Classification of synonyms, acronyms & long forms
- Classification reflects semantic significance

Name	Description	Examples
Modifier	Semantic-modifying tokens	receptor, inhibitor
Non-descriptive	Annotating tokens	fragment, precursor
Specifier	Numbers and Greek letters	1,VI, alpha, gamma
Common	Common English words	and, was, killer
Delimiter	Separator tokens	( ) , . ;
Standard	Standard tokens	TNF, BMP, IL

Table 1: Definition of token classes with differing semantic significance.

- „Standard“: IDs and anything else
- Classes are weighted for the search procedure

ProMiner

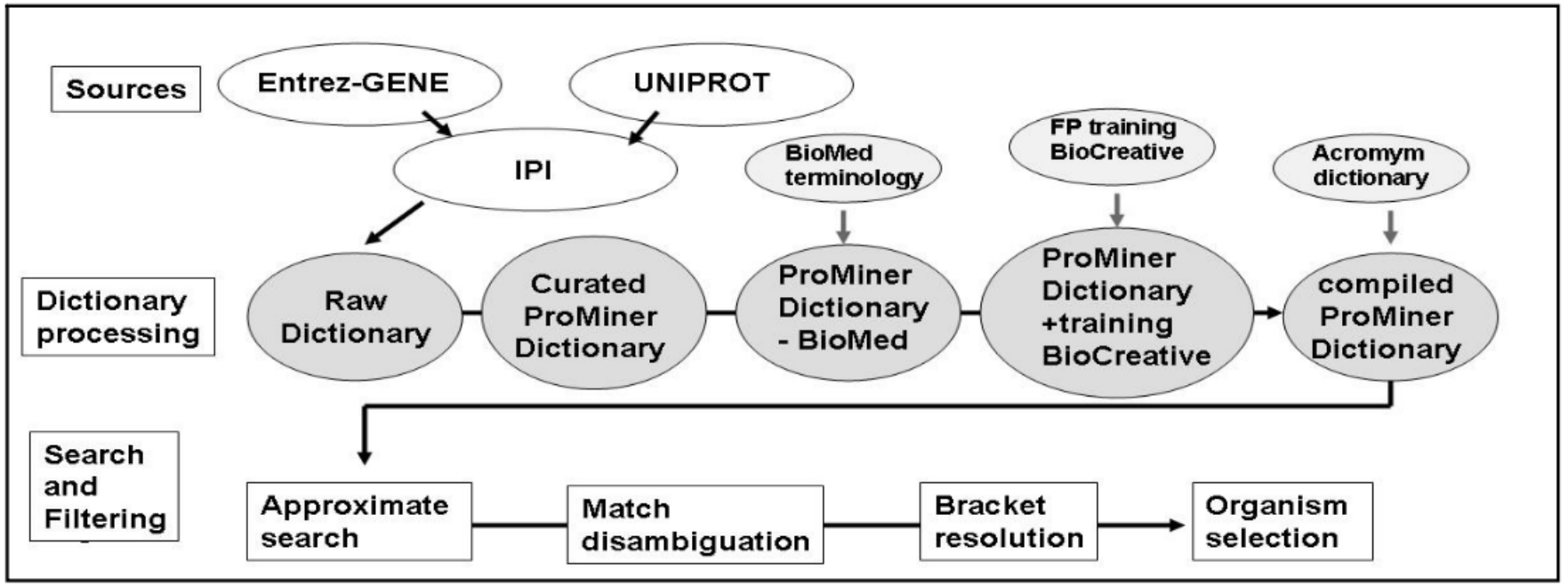


Figure 1: The ProMiner system used in BioCreAtIvE gene normalization

Approximate search

- Geared towards high sensitivity
- Variations in human terminology
  - permutations, insertions, deletions
  1. „Interleukin type 1 beta“ = „Interleukin-1 beta“
  2. „Interleukin-1 receptor“ ≠ „Interleukin 1“

Search procedure

- Token by token, with a set of candidates for the present position
- Candidate measurements
  - „boundary score“ is increased on mismatch, detects potential word boundaries
  - „acceptance score“ is a linear combination of
    - „match terms“
      - percentage of matched tokens per token class
    - „mismatch terms“
      - # of tokens in the text not found in the candidate

Match Terms

text	IL	type	1	
× cand. I	IL	-	1	
× cand. II	IL		1	receptor

Figure 1
First example of impact of token classes. Candidate synonym I is a correct synonym match, whereas candidate II is not. Appropriate weighting of tokens allows to detect the differences correctly.

- Exact matching: Ø
- Small weighting for ‚non-descriptive' tokens (-, type)
- High weighting for ‚modifiers' (receptor)

Mismatch Terms

text	Interleukin		1	receptor
× cand. I	Interleukin	-	1	
× cand. II	Interleukin		1	

Figure 2
Second example of impact of token classes. Both candidates are wrong matches because the significant token "receptor" is present in the text. Naive matching would accept both candidates.

- Naive matching would accept both
- Significant ‚modifier' „receptor" missing
- High mismatch weight for ‚modifiers'

- Weighting scheme
  - Based on a small benchmark
  - Penalizes deletion and insertion of ‚modifiers' heavily
  - Allows deletion and insertion of ‚non-descriptive' tokens
- Problems with the resulting set of synonyms
  - Overlapping matches → higher acceptance score („furrow" vs. „morphogenetic furrow")
  - Ambiguous synonyms

ProMiner

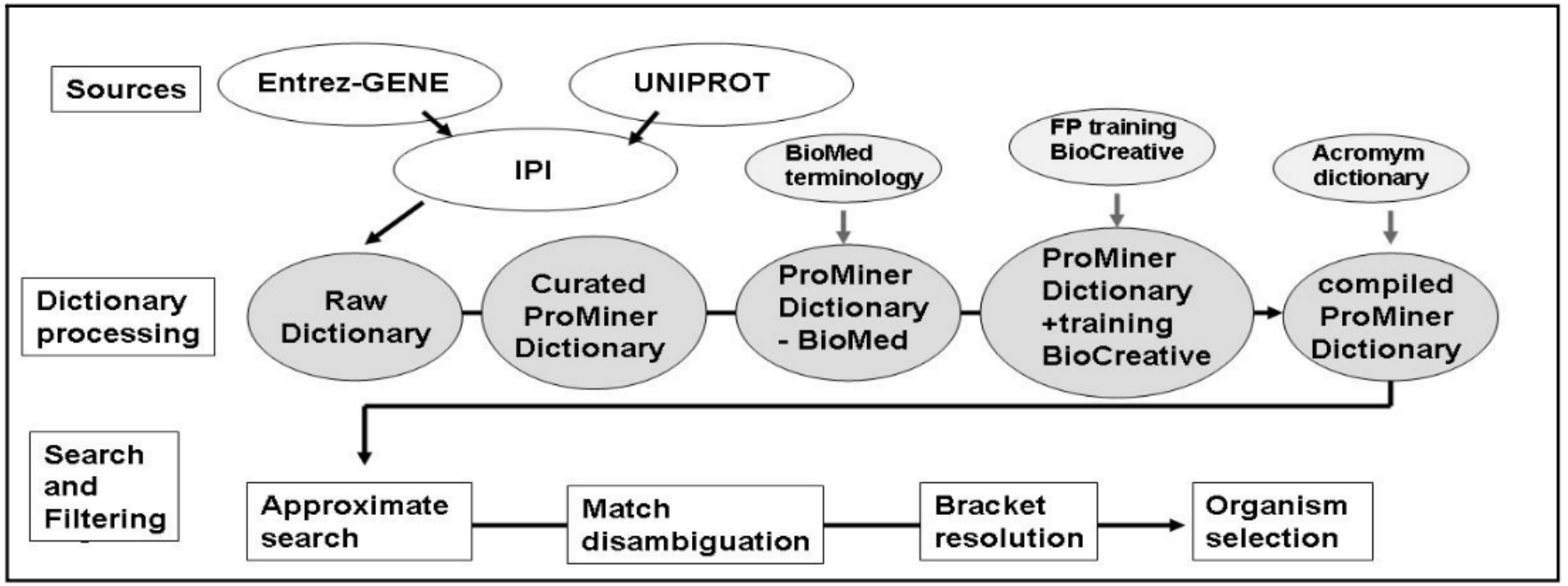


Figure 1: The ProMiner system used in BioCreAtIvE gene normalization

Match disambiguation

- Several potential IDs for a mention in the text
- ID with most additional synonym mentioned will be selected
  - No synonyms mentioned → ignore match
- User assigned synonymy threshold (D#)
  - # of synonyms > D# → ignore match

Bracket resolution

- Protein names can be split by acronyms in brackets („coenzyme A (HMG-CoA) synthase“)
- Combination of separate runs
  - Original text
  - Without brackets
  - Without bracketed expression
- Decision by ambiguity filter

Organism selection

- We only want abstracts about human genes
- Filter based on NCBI traxonomy database
- Simple organism name detection
  - Only irrelevant organisms → reject
  - Otherwise → accept

⇒ FPs if relevant and irrelevant organisms in text

Results in BioCreAtIvE II

- D1 (no ambiguity)
  - F-measure of 0.799
  - 3rd in BioCreative II
- D1 with original dictionary
  - Precision: 0.833 → 0.809
  - F-measure of 0.792
- D1 with organism detection
  - Precision: 0.833 → 0.835
  - Recall: 0.768 → 0.730
  - F-measure of 0.779
- Effect of bracket resolution unreproducable on the test set.

Rule-based approach

LMU München

Rule-based approach

- Gene name detection
  - Matching with BioCreAtIvE I systems
    - ProMiner (approximate matching)
    - Exact text matching
      - Simple, but close to the best results
      - No disambiguation
      - Large synonym lists (spelling variants)
    - Results are combined (CS)
- Post-matching (focus)
  - **Extended rule-based postfilter (RF)**
  - Abbreviation resolution
  - Disambiguation

Gene name detection

- Dictionary generation
  - Data from Entrez Gene, SWISSPROT and HUGO
  - Tuned towards Recall (two character synonyms)
  - ⇒ 32,969 genes with 587,250 synonyms (original dictionary: 168,805)

Rule-based postfilter (RF)

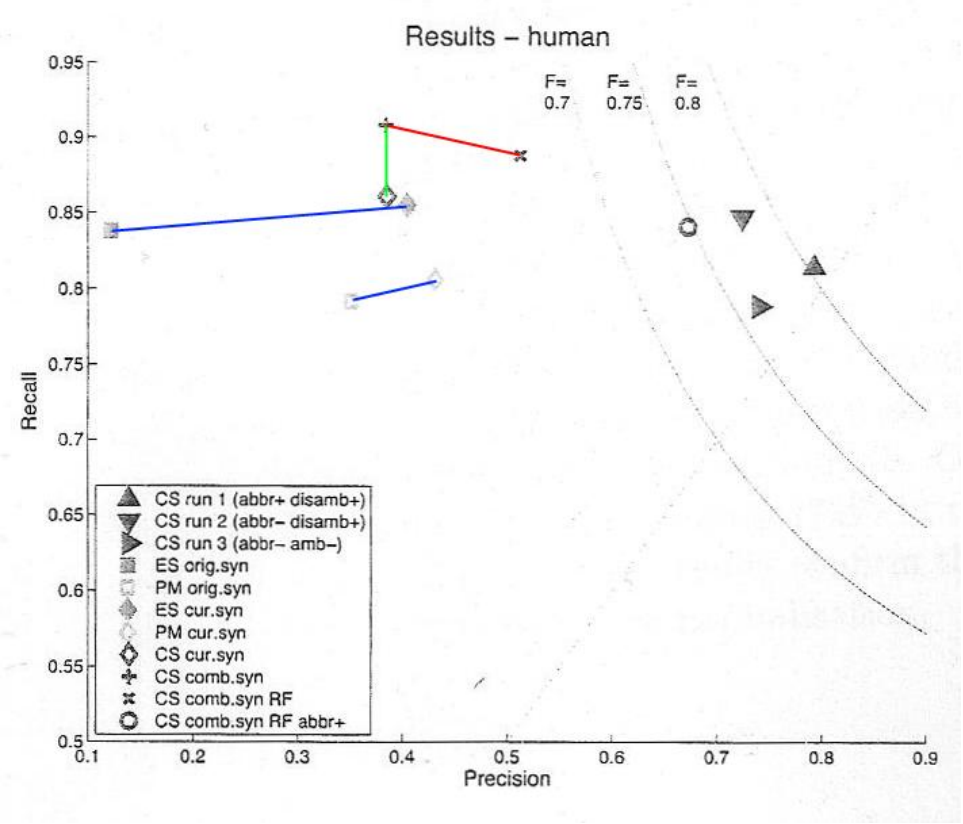
- Extended rule set
  - Unspecific words nearby (region, cell, family, ...)
  - Chromosome names („6p21.3“) followed by chromosome, region, band, ...
  - Chemical elements
  - Amino acid three-letter codes
  - Resolution of enumerations ending on Roman or Arabic numbers
    - ”IL-1 to IL-7”
  - …

Abbreviations & Ambiguity

- Special abbreviation dictionary
  - Collection of abbreviations and long forms
    - Combined with non-gene concepts of UMLS
    - Removal of long forms similar to dictionary synonyms
- Disambiguation using cosine similarity
  - NP chunks in the abstract
  - Synonyms of possible identfiers

  ⇒ Best rated synonym (if unique)

Results



Parameters	R	P	F
1: $abbr_+$ $disamb_+$	0.815	0.792	0.804
2: $abbr_-$ $disamb_+$	0.847	0.723	0.780
3: $abbr_-$ amb_-	0.789	0.739	0.763

- Organizers' dict.: P low
- Curated dict.: P much higher
- Own dict.: R higher
- Rule Filter: P higher
- abbr & dis: P, R and F higher

Conclusion

- Better F-measure than ProMiner (0.804 vs. 0.799)
- 2nd in BioCreative II
- Dictionary quality is essential
- Relies solely on dictionary information
  - No need for annotated training data
  - Yet competitive

BioTagger

Georgetown University Medical Center

BioTagger

- Based on Machine Learning
- Gene Mention Task
  - Dictionary from BioThesaurus and Metathesaurus
  - ML component with CRF(conditional random field)
    - Incorporates POS information (GENIA tagger)
  - Post-processing (abbreviations, parenthesis)
  - F-m of 0.859 (2nd quartile of 21 teams)
- Gene Normalization Task

Dictionary-lookup

- Synonym dictionary based on BioThesaurus and HUGO
- Search yields a list of pairs (Phrase, EGID)
- Enumeration expansion
  - ”HAP2-4”, ”HAP2/4”, ”HAP2, 3, 4”
  - Separate searches for ”HAP2” and ”HAP4”

Machine learning

- Feature extraction for each pair (Phrase, EGID)
  - Entity – Phrase detetected by GM module?
  - Exact match?
  - Ambiguity – number of EGIDs associated to Phrase
  - Number of references to EGID in the abstract
  - Primary or Synonym?
  - FP rate of the pair on noisy training data
  - Frequency of Phrase and EGID
  - Numbers, Greek letters?
  - Mixed case?
  - Punctuation or space nearby?
  - ….

- Fixed set of features for each pairs
  - Most standard ML algorithms can be used
- ML with Weka (JAVA ML package)
  - Cross validation of all algorithms
  - ”Bagging on Decision Tree” performed best
- Positive/Negative classification of pairs

Similarity-based mapping

- Problems with MWE synonyms
  - Deletions, insertions, permutations
- Simple solution
  - If > 90% of the words in a synonym name are found in the detected phrase, it will be normalized to the corresponding EGID.

Results

- 3 runs with different dictionaries
  1. Combination of 2nd and 3rd (how ?)
  2. Without frequent common English words
     Without names that resulted only in FP on „noisy“ test data
  3. Raw dictionary

Table 1: Gene mention (GM) and gene normalization (GN) results.

	Precision (Quartile)	Recall (Quartile)	F-Measure (Quartile)
GM-Run1	0.857 (2)	0.848 (2)	0.853 (2)
GM-Run2	0.834 (3)	0.880 (1)	0.856 (2)
GM-Run3	0.827 (3)	0.893 (1)	0.859 (2)
GN-Run1	0.743	0.824	0.781 (1)
GN-Run2	0.764	0.792	0.778 (1)
GN-Run3	0.790	0.769	0.779 (1)

- Dictionary hardly influences F-score, but Recall can be increased.
- Appropriate ML task works with standard dictionary
- 5th in BioCreative II

Conclusion on BioTagger

- Rich feature list in ML, but contribution of individual features is unclear.
- Main types of errors
  - Boundary detection errors
    ”v-rasHa retrovirus” instead of ”v-rasHa”
  - Ambiguity of short forms
  - FPs by non-specific mentions
    ”mouse genomic sequence”
- System is based on annotated corpora, which are expensive to obtain.

Me and my friends

“Tell me who your friends are, and I will tell you who you are.”

TU Dresden

Transinsight GmbH

Me and my friends

- Relies on semantically related information for ambiguity resolution
- Aspects that describe a gene
  - Localisation on a chromosomal band
  - Membership in a gene family
  - Molecular function
  - Mutations cause diseases
  - ...
- Whenever a gene is discussed, some of these aspects will be mentioned as well.

Methods

- Dictionary creation
- Named entity recognition
- FN detection
- Normalization
  - Reduction of ambiguity
  - Disambiguation of remaining terms and IDs

Finding FNs of the NER

- For each possible ID
  - Create a set of representative texts (noisy data, Entrez Gene Summary)
  - Turn representatives into feature vectors with **tf·idf** feature weights
- Filter the 100 most similar texts to the current abstract (cosine distance)

  ⇒ Get the IDs mentioned in these abstracts
  - Select IDs that share a synonym with the candidate name (approx. search)

Reduction of ambiguity

- Goal: detect FPs of the recognition module
- For every name mentioned
  - Create a **tf·idf** score
    (term frequency · inverse document frequency)
  - Low tf·idf score → drop (likely FP annotation)

Disambiguation of remaining IDs

- Comparison of each gene's (ID) context with the current text
- External knowledge on genes
  - **Entrez Gene**: summaries, GO terms
  - **UniProt**: gene functions, GO terms
  - **Gene Ontology Annotation**: GO terms
- Entrez Gene and UniProt
  - Calculate overlap of current text with each ID's annotation (token based)

  ⇒ 2 likelihoods

- Similarity based on GO terms
  - Find GO terms in the current text (using GoPubMed)
  - Find GO terms in the annotation of the ID (in Entrez Gene, UniProt and GOA)
  - For all possible pairs from these two sets
    - Compute distance in the ontology tree
    - Combine distance of all pairs

  ⇒ 3 likelihoods (one for each knowledge base)

- Combine all 5 likelihoods for each gene

⇒ ID with highest probability (threshold)

Results

Description of the submitted run	Precision	Recall	F1 (in %)	TP	FP	FN
NER with extended masterlist, FP+FN filter, disambiguation	78.9	83.3	81.0	654	175	131
NER with extended masterlist, FP filter, no disambiguation	49.6	87.5	63.3	687	699	98
NER with unextended masterlist, FP filter, disambiguation	70.7	72.5	71.6	569	236	216

- F-measure of 0.81
- 1st in BioCreative II

- Effect of FN detection cannot be determined (different conditions)

Conclusion on GN in BioCreAtIvE II

- Progress since BioCreAtIvE I in 2004
  - 9 teams achieved F ≥ 0.75
  - More participants (8 → 20)
  - Emergence of reusable components
- GN task still quite artificial
- Voting system of all teams could improve results (F-m > 0.83)
- Interdisciplinary approaches (ML, NLP, IR, biology, informatics)

Thanks for your attention!

References

- A. Morgan and L. Hirschmann, *Overview of BioCreative II Gene Normalization.*
  Proceedings of the Second BioCreative Challenge Evaluation Workshop. 17-27.
- D. Hanisch, K. Fundel, H.T. Mevissen, R. Zimmer, and J. Fluck (2005), *ProMiner: rule-based protein and gene entity recognition.*
  BMC Bioinformatics. 6(Suppl 1): S14.
- J. Fluck, H. Mevissen, H. Dach, M. Oster and M. Hofmann-Apitius (2006), *ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries.*
  Proceedings of the Second BioCreative Challenge Evaluation Workshop. 149-151.
- K. Fundel, D. Güttler, R. Zimmer and J. Apostolakis (2005), *A simple approach for protein name identification: prospects and limits.*
  BMC Bioinformatics. 6(Suppl 1): S15.
- D. Hanisch, J. Fluck, H. Mevissen and R. Zimmer (2003), *Playing Biology's Name Game: Identifying Protein Names in Scientific Text.*
  Pacific Symposium on Biocomputing, 8:403-414.
- K. Fundel and R. Zimmer, *Human Gene Normalization by an Integrated Approach including Abbreviation Resolution and Disambiguation.*
  Proceedings of the Second BioCreative Challenge Evaluation Workshop. 153-155
- Hanisch D, Fundel K, Mevissen H-T, Zimmer R and Fluck J, *ProMiner: Organism-specific protein name detection using approximate string matching.*
  BMC Bioinformatics 2005, 6(Suppl 1):S14.
- K. Fundel, D. Guettler, R. Zimmer, and J. Apostolakis (2004). *Exact versus approximate string matching for protein name identification.*
  Proceedings of the BioCreative Challenge Evaluation Workshop 2004.
- J. Hakenberg, L. Royer, C. Plake, H. Strobelt, and M. Schroeder (2007), *Me and my friends: gene mention normalization with background knowledge.*
  Proceedings of the Second BioCreative Challenge Evaluation Workshop, 141-144.
- H. Liu, M. Torii, ZZ Hu, and C. Wu (2007), *Gene Mention and Gene Normalization Based on Machine Learning and Online Resources.*
  Proceeding of the Second BioCreative Challenge Evaluation Workshop, 135-140.
- Liu H, Wu C and Friedman C. (2004), *BioTagger: a biological entity tagging system.*
  Proceedings of the biocreative challenge evaluation workshop, 2004, Grenada.