Universität Bielefeld
Faculty of Technology – Semantic Computing Group

Towards Gene Recognition from Rare and Ambiguous Abbreviations using a Filtering Approach

Matthias Hartung*, Roman Klinger*, Matthias Zwick‡ and Philipp Cimiano*

* Semantic Computing Group, Cognitive Interaction Technology Center of Excellence (CIT-EC), Bielefeld, Germany
‡ Research Networking, Boehringer Ingelheim Pharma GmbH, Biberach, Germany

Summary

- Rare and ambiguous gene name abbreviations are infrequently represented in common benchmarking resources.
- We release a manually annotated corpus [3] that enables a more thorough investigation of the phenomenon.
- Our combination of a CRF and an SVM outperfoms existing gene recognition tools by 9 points F_1 score in a cross-entity evaluation on this corpus.

Motivation

- Reference to genes/proteins by abbreviations is ubiquitous in biomedical publications.
- Thesaurus-based search engines: automatic query expansion by all known synonyms and abbrevations
- For rare and ambiguous gene/protein abbreviations: high proportion of false positives
  ⇒ hard challenge for text mining in pharmaceutical research

Gene Name Abbreviations Investigated

Synonym	Other name(s)	Other meaning(s)
SAH	acyl-CoA synthetase medium-chain family member 3; ACSM3	subarachnoid hemorrhage; S-Adenosyl-L-homocysteine hydrolase
MOX	monooxygenase, DBH-like 1	moxifloxacin; methylparaoxon
PLS	POLARIS	partial least squares; primary lateral sclerosis
CLU	clusterin; CLI	covalent linkage unit
CLI	clusterin; CLU	clindamycin
HF	complement factor H; CFH	high frequency; heart failure; Hartree-Fock
AHR	aryl hydrocarbon receptor; bHLHe76	airway hyperreactivity
COPD	archain 1; ARCN1; coatomer protein complex, subunit delta	Chronic Obstructive Pulmonary Disease

Corpus Analysis: Standard Data Sets

Protein	MEDLINE # Tokens (abs.)	MEDLINE # Tokens (per 1M)	MEDLINE % tagged	BioCreative2 # Tokens (abs.)	BioCreative2 # Tokens (per 1M)	BioCreative2 % of genes	GENIA # Tokens (abs.)	GENIA # Tokens (per 1M)	GENIA % of genes
SAH	30019	8.58	6.1 %	2	3.33	0 %	0	0	
MOX	16007	4.57	13.1 %	0	0		0	0	
PLS	11918	3.41	25.9 %	0	0		0	0	
CLU	1077	0.31	29.1 %	0	0		0	0	
CLI	1957	0.56	4.8 %	4	6.67	0 %	0	0	
HF	42563	12.16	7.9 %	8	13.33	62.5 %	4	8.89	0 %
AHR	21525	6.15	75.7 %	12	20.00	91.7 %	0	0	
COPD	44125	12.61	0.6 %	6	10.00	0 %	0	0	

Our Corpus: Properties and Statistics

- random sample of 100 MEDLINE abstracts per gene/protein with ≥ 1 occurrence of abbreviation of interest
- all abbreviations of interest manually annotated by two authors
- all tokens in abstract manually annotated by one domain expert

Protein	Pos. Inst.	Neg. Inst.	Total
SAH	5	349	354
MOX	62	221	283
PLS	1	206	207
CLU	235	30	265
CLI	11	211	222
HF	2	353	355
AHR	53	80	133
COPD	0	250	250

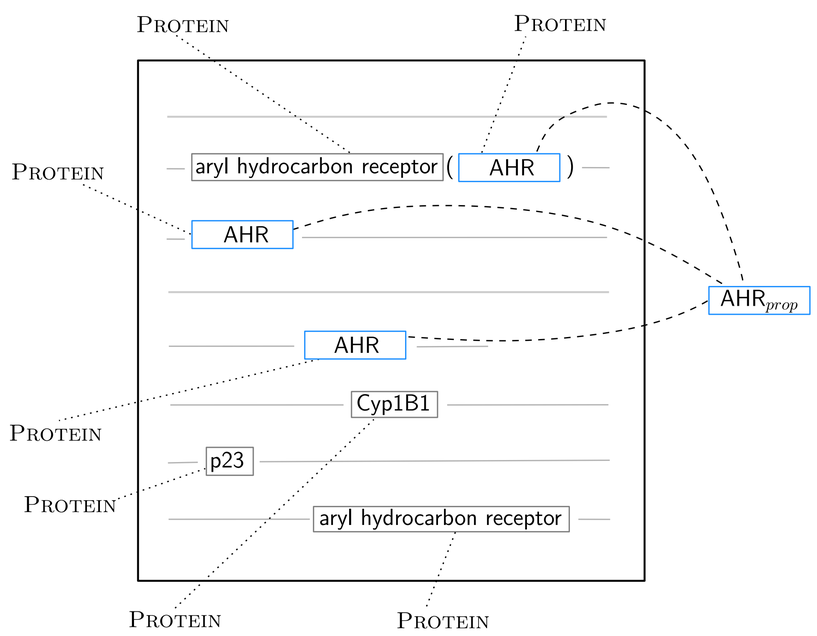
Features for Classification

Local Information:

- context: 6+4 tokens
- abbreviation potential
- tagger$_{local}$

Global Information:

- all unigrams in abstract
- tagger$_{global}$
- longform in abstract
- feature propagation



Cross-Entity Evaluation

- train on abstracts for 7/8 proteins
- test on abstracts for 1/8 proteins

Overall Performance

Setting	P	R	F_1
SVM	0.81	0.45	0.58
CRF∩SVM	**0.99**	0.26	0.41
CRF→SVM	0.83	0.49	0.62
CRF→SVM+FS	0.97	**0.74**	**0.84**
GNAT [2]	0.73	0.45	0.56
CRF [4]	0.55	0.43	0.48
AcroTagger [1]	0.92	0.63	0.75
longform	0.98	0.65	0.78
lex	0.18	1.00	0.32

Conclusions

- best-performing configuration: longform, context, tagger$_{global}$, propagation
- no gene-specific features, easily extensible approach
- gene abbreviations deserve special treatment beyond standard approaches
- future work: hybrid approach, using longform as decision rule

References

[1] S. Gaudan, H. Kirsch, and D. Rebholz-Schuhmann. Resolving abbreviations to their senses in Medline. *Bioinformatics*, 21(18):3658–3664, 2005.

[2] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, and G. Gonzalez. Inter-species normalization of gene mentions with GNAT. *Bioinformatics*, 24(16):i126–i132, Aug 2008.

[3] M. Hartung and M. Zwick. A Corpus for the Development of Gene/Protein Recognition from Rare and Ambiguous Abbreviations, 2014. Bielefeld University. doi:10.4119/unibi/2673424.

[4] R. Klinger, C. M. Friedrich, J. Fluck, and M. Hofmann-Apitius. Named Entity Recognition with Combinations of Conditional Random Fields. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, Madrid, Spain, April 2007.

Contact: Matthias Hartung, mhartung@cit-ec.uni-bielefeld.de

→ www.uni-bielefeld.de