

Term-based literature mining from biomedical texts

S. Ananiadou, G. Nenadic, I. Spasic
Computer Science, University of Salford, UK

D. Schuhmann
Lion Bioscience, Heidelberg, Germany

Outline

- ◆ term-based framework for literature mining:
 - automatic term recognition
 - term variation handling and acronym acquisition
 - automatic discovery of term similarities and term clustering
- ◆ ATRACT
- ◆ experiments & results
- ◆ conclusions

Automatic term recognition

- ◆ C/NC-value method for extraction of **multiword** terms
- ◆ hybrid approach incorporating:
 - linguistics
 - statistics
 - contextual information
- ◆ assigns weights to candidate terms and ranks them

Automatic term recognition

- ◆ linguistic information (extracting term candidates)
 - term formation patterns
- ◆ statistics (ranking term candidates)
 - frequency of occurrence
 - term length
 - frequency of nested terms
- ◆ contextual information (re-ranking term candidates)
 - co-occurrence with significant context words

Term variation handling

- ◆ naming conventions in biology/biomedicine are highly non-standardised
- ◆ terms should be mono-referential, but in practice we deal with:
 - **term ambiguities**: same term corresponds to many concepts
 - **term variants**: many terms leading to the same concept
- ◆ term variations handling: support for systematic knowledge acquisition

Sources of term variation

- ◆ orthographical
all-trans-retinoic acid vs. all trans retinoic acid
- ◆ morphological
biochemical study vs. biochemical studies
- ◆ syntactic
human clones vs. clones of humans
- ◆ acronyms
NF-kappa B vs. nuclear factor-kappa B
- ◆ lexico-semantic
ophthalmological surgery vs. eye surgery

Handling term variation

- based on normalisation of term candidates as an integral part of the ATR process
- example:

human cancers
cancer in humans
human's cancer
human carcinoma

} → human cancer

- term normalisation is performed prior to statistical analysis: a term candidate comprises all of its variations

Acronym acquisition

- frequently occurring in scientific texts
- method based on:
 - syntactic features of **acronym definitions**
... NRs (nuclear receptors) ...
 - morphological features** of acronym constituents (combining affixes: prefixes/suffixes)
... chloramphenicol **acetyl**(transferase (CAT) ...
 - clustering **acronym variants**
nuclear factor kappa B: NF-kappaB, NF kappa B, NF(kappa)B, kappaB, NFKB factor, NF-KB, NF kB

Acronym acquisition (examples)

acronyms	expanded forms	normalised term
RAR alpha RAR-alpha RARA RARa	retinoic acid receptor alpha retinoic acid receptor alpha retinoic acid receptor alpha	retinoic acid receptor alpha
APL	acute promyelocytic leukaemia acute promyelocytic leukemia	acute promyelocytic leukemia
9-C-RA 9CRA	9-cis-retinoic acid 9-cis retinoic acid	9-cis retinoic acid
RAR RARS	retinoic acid receptor retinoic acid receptors	retinoic acid receptor

Conflating term variants

term	C/NC-value
retinoic acid receptor retinoic acid receptor retinoic acid receptors RAR, RARS	6.33
nuclear receptor nuclear receptor nuclear receptors NR, NRS	6.00
all-trans retinoic acid all trans retinoic acid all-trans-retinoic acids ATRA, at-RA	4.25

Discovering term similarities

- extracted terms need to be associated with other terms
- classification/clustering of semantically similar terms
- term similarities are mined from literature
- hybrid method combining:
 - contextual similarity
 - syntactic similarity
 - lexical similarity

Contextual similarity (CS)

- automatically identify most important context patterns (CP) terms tend to appear in
- although automatically discovered from corpus, CPs are domain-specific!
- context patterns consist of:
 - terms automatically recognised in text
 - significant sentence chunks, optionally instantiated (e.g. NP, V:bind, PREP:with)
- significance of CPs is estimated by statistical analysis (frequency, length, linear nestedness)

Lexical similarity (LS)

- based on sharing a head and/or modifier(s)

- examples:

progesterone receptor
oestrogen receptor

nuclear receptor
orphan nuclear receptor

- lexical similarity on its own is not sufficient

Bride of sevenless

Syntactic similarity (SS)

- parallel usage of terms within the same context (e.g. enumerations, coordination, etc.) indicates their functional similarity

... both Swiss 3T3 fibroblasts and human T-lymphoblasts ...

- sample patterns

Term such as (Term,)+ CONJ Term

both Term and Term

either Term or Term

neither Term nor Term

Hybrid similarity measure

- linear combination of CS, LS and SS:

$$CLS(t_p, t_2) = a CS(t_p, t_2) + b LS(t_p, t_2) + g SS(t_p, t_2)$$

- parameters are automatically learnt from similarities based on an existing ontology
- results: similar terms tend to share similar "friends"

Sample similarity values

	glucocorticoid receptor	estrogen receptor	D
steroid receptor	0.66	0.64	0.02
progesterone receptor	0.55	0.59	0.04
human estrogen receptor	0.28	0.37	0.09
retinoid X receptor	0.27	0.36	0.09
nuclear receptor	0.30	0.33	0.03
receptor complex	0.31	0.33	0.02
retinoic acid receptor	0.27	0.28	0.01
retinoid nuclear receptor	0.26	0.26	0.00

Term Clustering

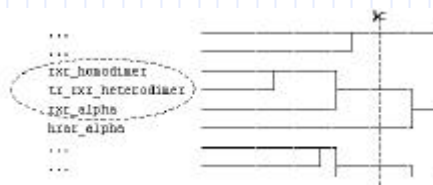
- hybrid similarity measure is used to produce a similarity matrix

- row: similarity vector for a specific term
- distances between vectors are used to establish clusters

- clustering methods and precision achieved:

- nearest neighbour: 63%
- Ward's method: 71%

Sample cluster



ATTRACT

- ATTRACT is an integrated terminology management system in the domain of molecular biology
- literature mining is carried out by integrating:
 - automatic term recognition
 - automatic term variants conflation
 - automatic term clustering

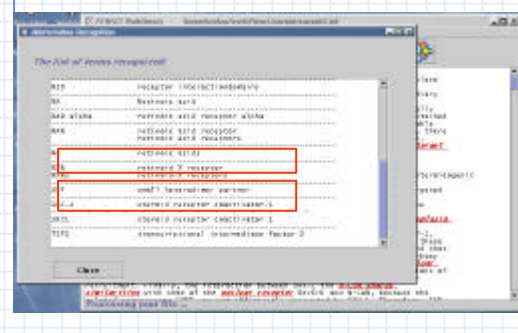
ATTRACT - term variants in text



ATTRACT - similar terms in text



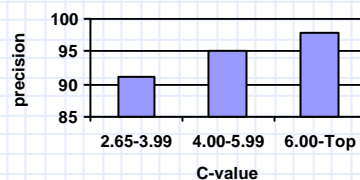
ATTRACT - acronym retrieval



Experiments

- domain: biomedicine
- tool: ATTRACT workbench
- source: MEDLINE database
- corpus: 2082 abstracts

Evaluation – term recognition



Evaluation – acronym acquisition

acronyms	corpus	
	2082 abstracts	50 abstracts
# of (distinct) acronyms recognised	1015	66
# of correct acronyms recognised	992	62
# of acronyms introduced	-	85
precision	97.73%	93.94%
recall	-	72.94%

Evaluation – term clustering

cardinality of a cluster	nearest neighbour			Ward's method		
	# of clusters	# of correct		# of clusters	# of correct	
		clusters	terms		clusters	terms
2	16	7(44%)	14	33	22(67%)	44
3	7	6(86%)	18	19	10(53%)	30
4	4	2(50%)	8	5	3(60%)	12
≥ 5	10	7(70%)	47	2	1(50%)	8
Total:	37	22(59%)	87(63%)	59	36(61%)	14(71%)

Conclusions

- ✦ literature mining based on effective management of terms and their variants
- ✦ term variants unification and normalisation - broader basis for IR and IE tasks
- ✦ hybrid term similarity measure: similar terms share most of their "friends"
- ✦ results - precision:
 - term recognition: 91-98%
 - acronym recognition: 94-99%
 - term clustering: 63-71%