Temporal Classifiers for Predicting the Expansion of Medical Subject Headings

George Tsatsaronis
Biotechnology Center, Technische Universität Dresden, Germany
george.tsatsaronis@biotec.tu-dresden.de
Iraklis Varlamis
Department of Informatics and Telematics, Harokopio University of Athens, Greece
varlamis@hua.gr
Nattiya Kanhabua
L3S Research Center, Leibniz Universität Hannover, Germany
kanhabua@l3s.de
Kjetil Nørvåg
Department of Computer and Information Science, Norwegian University of Science and Technology, Norway
Kjetil.Norvag@idi.ntnu.no

Annotation of biomedical data

- The amount of biomedical data increases exponentially
  - Scientific articles, nucleotide sequences, protein structures
- Ontology based annotation of data facilitates information indexing and retrieval
  - PubMed uses Medical Subject Headings (MeSH) to annotate Medline articles
  - GoPubMed uses Gene Ontology (GO) and UniProt resources
- Automatic annotation with ontology terms facilitates indexing
- … but who maintains the ontology?

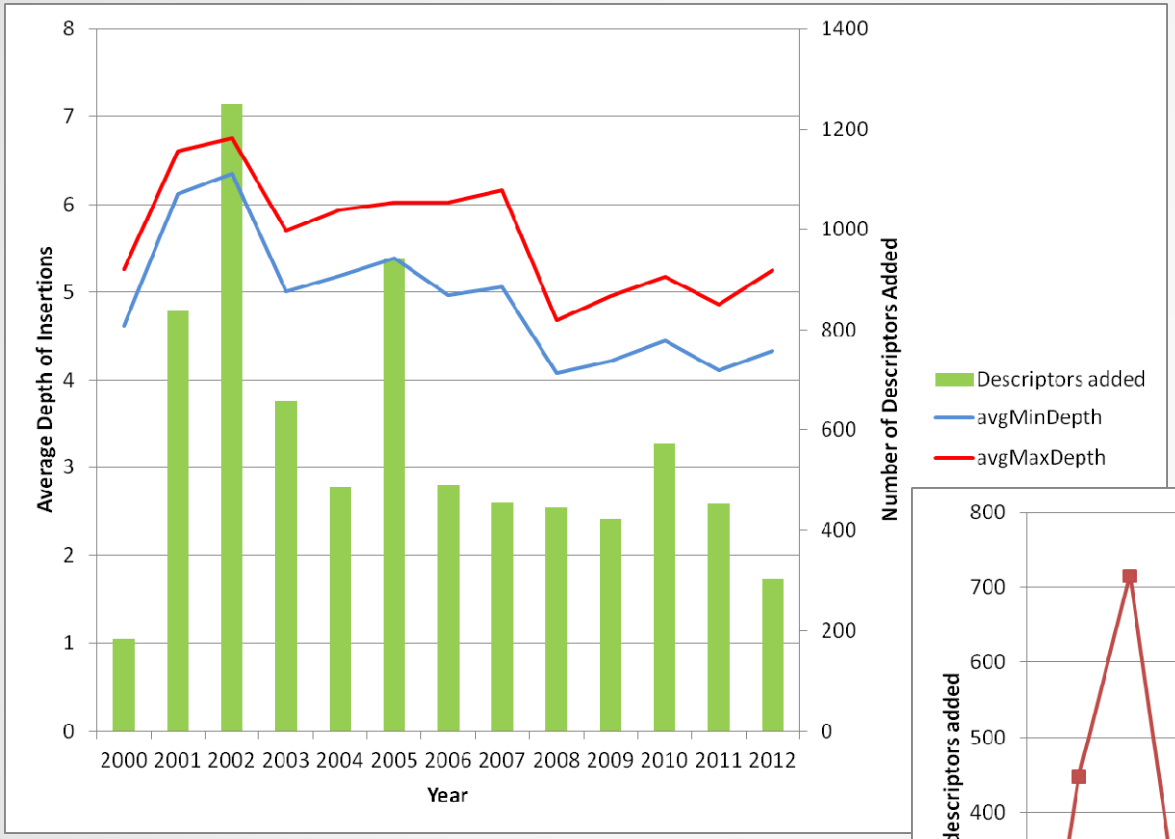
Ontology evolution

- Biomedical literature introduces new terms, which should be incorporated to the ontology
- Open questions in ontology evolution
  - Which terms to include in the ontology?
  - Where to place this terms?
- We present a methodology for updating MeSH hierarchy
- We predict which MeSH headings may be expanded in the near future

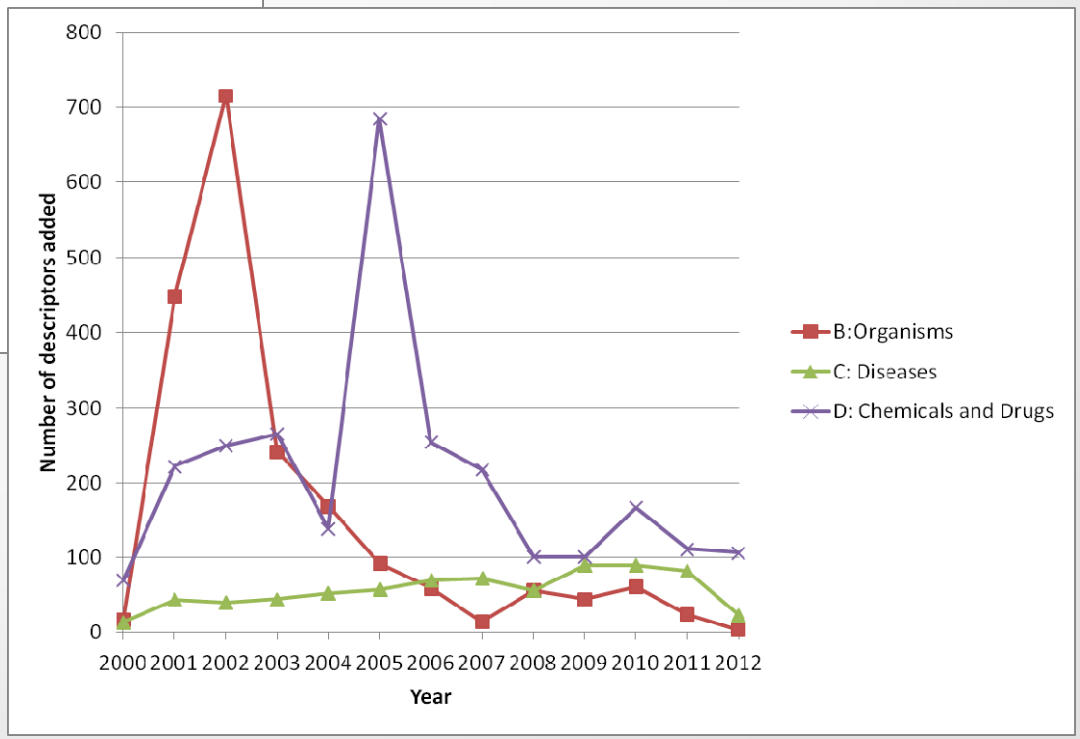
Medical Subject Headings (MeSH)

- Hierarchy of terms maintained by the United States National Library of Medicine
- MeSH includes three types of data:
  - descriptors (subject headings): 26,853 terms in 16 trees,
  - qualifiers (subheadings): 80 terms that narrow the descriptors' topics
  - supplementary concept records: 214,000 terms that mainly describe chemical substances and are linked to the respective descriptors
- Our classifiers use both static and temporal features of the descriptors to predict which of them will be expanded in the forthcoming MeSH releases

MeSH statistics



- 576 new descriptors are added on average per year
- changes occur on average between depths 4 and 6



Trees B, C, D
- number approximately 17, 000 MeSH headings (64% of the MeSH hierarchy)
- contain approximately 68% of all new MeSH headings additions since 1999 (5388 descriptors added)

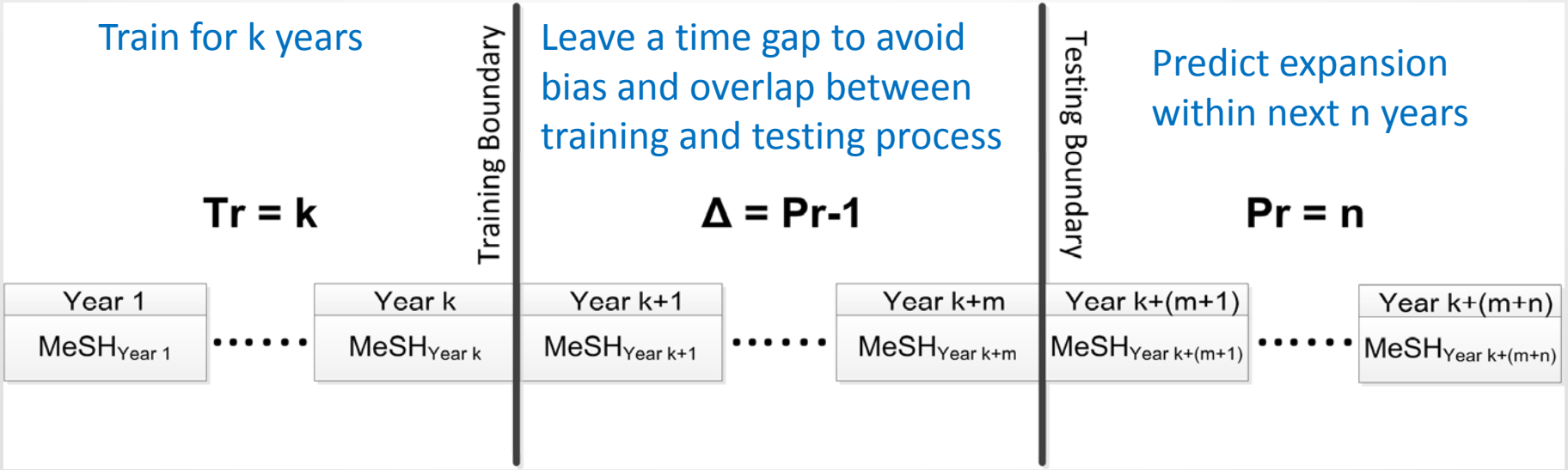
Related projects

- Predict the extension of the *Gene Ontology* (Pesquita and Couto, 2012)
  - predict areas of GO that will undergo extension in a future version
  - structural, annotation, citation and hybrid features are employed
  - only the latest version of the ontology is employed
  - Our method all the previous versions of the ontology to define temporal features
- DOG4DAG: automated sibling generation for *MeSH* terms (Fabian, Wächter and Schroeder, 2011)
  - Siblings are extracted from the terms' context.
  - Simple co-occurrence and textual patterns (`A *such as* B, C, D`) are employed
  - our method can be used to find the terms that will be expanded

MeSH expansion as a classification problem

- I is a MeSH descriptor
- We compute the values of all features of I up to year t using all MeSH snapshots from 1999 to t: $I_t = [X_1, ..., X_N]$.
- If I is to be expanded in the next n years then $C=1$, else $C=0$

$$M : I \times C \rightarrow 0, 1$$



Features

Category	Feature	Name	Description
Structural	X_1	*minDepth*	Minimum depth I appears
	X_2	*maxDepth*	Maximum depth I appears
	X_3	*siblings*	# *MeSH* heading siblings to I
	X_4	*direct children*	# *MeSH* heading direct children of I
	X_5	*all children*	# *MeSH* heading descendants of I
Citation	X_6	*PubMed results*	# *PubMed* results with I as query
	X_7	*direct children results*	# *PubMed* results with I's children as query
	X_8	*all children results*	# *PubMed* results with I's descendants as query
Annotation	X_9	*PubMed annotations*	# major/minor *PubMed* annotations with I
	X_{10}	*direct children annotations*	# major/minor *PubMed* annotations with I's children
	X_{11}	*all children annotations*	# major/minor *PubMed* annotations with I's descendants
Hybrid	X_{12}	*annRatioAll*	$\frac{\textit{all children annotations}}{\textit{all children}}$
	X_{13}	*annRatioDir*	$\frac{\textit{direct children annotations}}{\textit{direct children}}$
	X_{14}	*resRatioAll*	$\frac{\textit{all children results}}{\textit{all children}}$
	X_{15}	*resRatioDir*	$\frac{\textit{direct children results}}{\textit{direct children}}$
	X_{16}	*annRatioResults*	$\frac{\textit{PubMed annotations}}{\textit{PubMed Results}}$
Temporal	X_i'	*temporal* X_i	$\frac{X_{i,y_n} - X_{i,y_{n-1}}}{X_{i,y_n}}$

Experimental setup

- Ontology: We are using all the MeSH releases from 1999 until 2011 (inclusive)
- Corpus: 22 million articles indexed in PubMed until 31/12/2011
- Indexed in Lucene: titles, abstracts, years, MeSH major and minor annotations of all articles in the corpus; all MeSH releases
- Classifier: cost sensitive classification (MetaCost classifier with pre-constructed cost matrices) on-top of Random Forests (in Weka)
- Different classifiers for different MeSH trees

Instance files

We create 12 Weka files (1/y) for each MeSH tree and then merge instances

Tr	*PR*= 1($\Delta = 0$)	*PR*= 2($\Delta = 1$)	*PR*= 3($\Delta = 2$)	*PR*= 4($\Delta = 3$)	*PR*= 5($\Delta = 4$)	*PR*= 6($\Delta = 5$)
1	11	9	7	5	3	1
2	10	8	6	4	2	–
3	9	7	5	3	1	–
4	8	6	4	2	–	–
5	7	5	3	1	–	–
6	6	4	2	–	–	–
7	5	3	1	–	–	–
8	4	2	–	–	–	–
9	3	1	–	–	–	–
10	2	–	–	–	–	–

4 years for training
2 years prediction
1 year gap

6 evaluation tests

Evaluation of results

- Micro-averaged Precision (P), Recall (P) and F-Measure (F) for the positive class (C = 1)
- The best results for each Pr value are reported in bold (usually when all instances from the previous periods are employed)

Tr	*PR*= 1($\Delta = 0$)			*PR*= 2($\Delta = 1$)			*PR*= 3($\Delta = 2$)			*PR*= 4($\Delta = 3$)			*PR*= 5($\Delta = 4$)			*PR*= 6($\Delta = 5$)		
-	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
1	16.7	0.5	1.0	11.5	29.1	16.5	38.1	27.1	31.7	14.9	48.8	22.8	11.9	62.3	19.9	**12.1**	**62.8**	**20.3**
2	4.8	12.0	5.6	10.8	37.6	16.8	20.8	37.6	26.7	32.9	33.1	33.0	12.9	65.0	21.5	-	-	-
3	3.2	13.5	4.7	8.7	41.0	14.3	30.8	30.7	30.7	36.7	33.6	35.1	**16.4**	**35.0**	**22.3**	-	-	-
4	6.5	13.2	7.2	12.9	30.2	18.1	29.1	28.6	28.8	43.5	40.4	41.9	-	-	-	-	-	-
5	11.4	10.3	9.3	13.1	27.7	17.7	28.1	33.5	30.6	**64.9**	**53.6**	**58.7**	-	-	-	-	-	-
6	8.0	9.8	8.5	20.1	30.6	24.3	59.3	40.0	47.8	-	-	-	-	-	-	-	-	-
7	12.5	5.6	7.7	16.1	20.7	18.1	**88.2**	**50.0**	**63.8**	-	-	-	-	-	-	-	-	-
8	14.3	4.8	7.1	35.3	17.9	23.8	-	-	-	-	-	-	-	-	-	-	-	-
9	35.0	5.4	9.0	**28.4**	**28.1**	**28.3**	-	-	-	-	-	-	-	-	-	-	-	-
10	**42.9**	**14.3**	**21.4**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Results for *MeSH Tree B (Organisms)*

Results on Diseases

Tr	*PR*= 1($\Delta = 0$)			*PR*= 2($\Delta = 1$)			*PR*= 3($\Delta = 2$)			*PR*= 4($\Delta = 3$)			*PR*= 5($\Delta = 4$)			*PR*= 6($\Delta = 5$)		
-	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
1	16.4	1.2	2.3	37.8	7.6	12.7	40.4	16.3	23.2	44.4	30.2	35.9	36.3	22.9	28.1	**37**	**36.9**	**36.9**
2	31.3	1.4	2.8	39.7	6.2	10.7	45.6	16.5	24.2	45.1	36.4	40.3	41.3	30.3	35	-	-	-
3	26.5	2	3.8	49.1	7.8	13.5	45.4	21.7	29.4	46.2	48.5	47.3	**40.4**	**39.8**	**40.1**	-	-	-
4	36	1.5	2.9	34.7	6.9	11.5	42.7	28.3	34	65.8	43.8	52.6	-	-	-	-	-	-
5	25	1.9	3.5	38.3	12.8	19.3	51.1	34.2	41	**79.6**	**57**	**66.4**	-	-	-	-	-	-
6	26.3	3.2	5.7	35.7	12.9	18.9	67.7	34.7	45.9	-	-	-	-	-	-	-	-	-
7	32.4	6.1	10.2	40.5	16.7	23.7	**81.9**	**45.4**	**58.4**	-	-	-	-	-	-	-	-	-
8	36.1	4.1	7.3	48.4	23.7	31.8	-	-	-	-	-	-	-	-	-	-	-	-
9	37.5	5.8	10.1	**53.1**	**41**	**46.2**	-	-	-	-	-	-	-	-	-	-	-	-
10	**47.4**	**9.3**	**10.9**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Results for *MeSH Tree C (Diseases)*

Results on Drugs

Tr	*PR*= 1($\Delta = 0$)			*PR*= 2($\Delta = 1$)			*PR*= 3($\Delta = 2$)			*PR*= 4($\Delta = 3$)			*PR*= 5($\Delta = 4$)			*PR*= 6($\Delta = 5$)		
-	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
1	54.5	5.2	9.5	46.4	11.8	18.8	36.5	26.8	30.9	44.1	39.1	41.4	32.4	35.6	33.9	**30.4**	**34.1**	**32.1**
2	53.9	5.8	10.5	48.2	12.6	20	33.5	33.9	33.7	50.1	36.7	42.3	27	38.4	31.7	-	-	-
3	33.3	5.9	10	36.7	15.5	21.8	47.1	28.4	35.5	59	38.5	46.6	**41.4**	**30.9**	**35.4**	-	-	-
4	51.3	8.5	14.5	43.8	17.9	25.5	39	37.8	38.4	52.9	45.9	49.2	-	-	-	-	-	-
5	22.3	17.3	19.5	37.4	21.7	27.5	53.8	32.7	40.7	**63.2**	**56.5**	**59.7**	-	-	-	-	-	-
6	**32**	**15.7**	**21.1**	41.6	25.9	31.9	49.1	36	41.6	-	-	-	-	-	-	-	-	-
7	23.9	16.3	19.4	30.5	31.5	31	**64.5**	**51.7**	**57.4**	-	-	-	-	-	-	-	-	-
8	31.2	14.1	19.4	57.7	22	31.8	-	-	-	-	-	-	-	-	-	-	-	-
9	53.6	8.3	14.4	**75.3**	**41.7**	**53.6**	-	-	-	-	-	-	-	-	-	-	-	-
10	34.8	11	16.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Results for *MeSH Tree D (Chemicals and Drugs)*

Conclusions from the evaluation

- More years for training (higher Tr) gives better results
- For Tree B: prediction performance reached precision of 88.9% with recall of 50% (F-Measure= 63.8%), in maximum three years from the testing year
- The prediction of expansions using Pr in the range [2, 4] is possible with satisfactory results, if >=5 training years are used
- Performance in predicting expansion for the next 1 year is poor
- Top-5 features: temporal siblings, temporal all children, temporal direct children, annRatioAll, all children results,
  - Temporal features aid significantly the prediction,
  - the use of the offered PubMed annotations, and the use of PubMed corpus are extremely beneficial.

General conclusions

- Our methodology
  - under conditions can predict the MeSH regions that will be expanded with a relatively high Precision, if sufficient number of training instances is provided, and a lengthier prediction span is given as a parameter
  - is a first step for automated ontology evolution, provided that it is augmented with a second step, which may also suggest specific new terms to be added below the MeSH headings that are predicted as positive

Next steps

- Find the terms that will extend the MeSH hierarchy
- Evaluate the methodology on ontologies (e.g. GO) or other domains (e.g. patents) with similar structure

Thank you!

Questions?

Iraklis Varlamis

varlamis@hua.gr