BioASQ Workshop on Conference and Labs of the Evaluation Forum (CLEF 2014)
Sheffield, UK
September 16-17, 2014

The Application of NCBI Learning-to-rank and Text Mining Tools in BioASQ 2014

Yuqing Mao, Chih-Hsuan Wei, Zhiyong Lu

National Center for Biotechnology Information
Bethesda, Maryland - USA

NCBI

NIH U.S. National Library of Medicine

BioASQ Task 2a

- Large-scale Online Biomedical Semantic Indexing
  - Provided with a set of newly published articles in PubMed
    - Automatically predict the most relevant MeSH terms for each articles
  - Evaluated by comparing the results to the gold standard curated by human indexers

Challenges in MeSH Indexing

- **Scale**: MeSH 2014 includes 27,000+ main headings (e.g. Humans, Parkinson Disease)
- **Complex cognitive task**: Consistency among human indexers is 48.2% for main heading assignment (Funk et al., 1983)
- **Time to index varies**: 25% of the citations are completed within 30 days of receipt, 50% within 60 days, and 75% within 90 days (Huang et al., 2010, 2011)

Huang, Neveol, & Lu (2011). Recommending MeSH terms for annotating biomedical articles, JAMIA
Huang & Lu (2010). Learning to annotate scientific publications, COLING

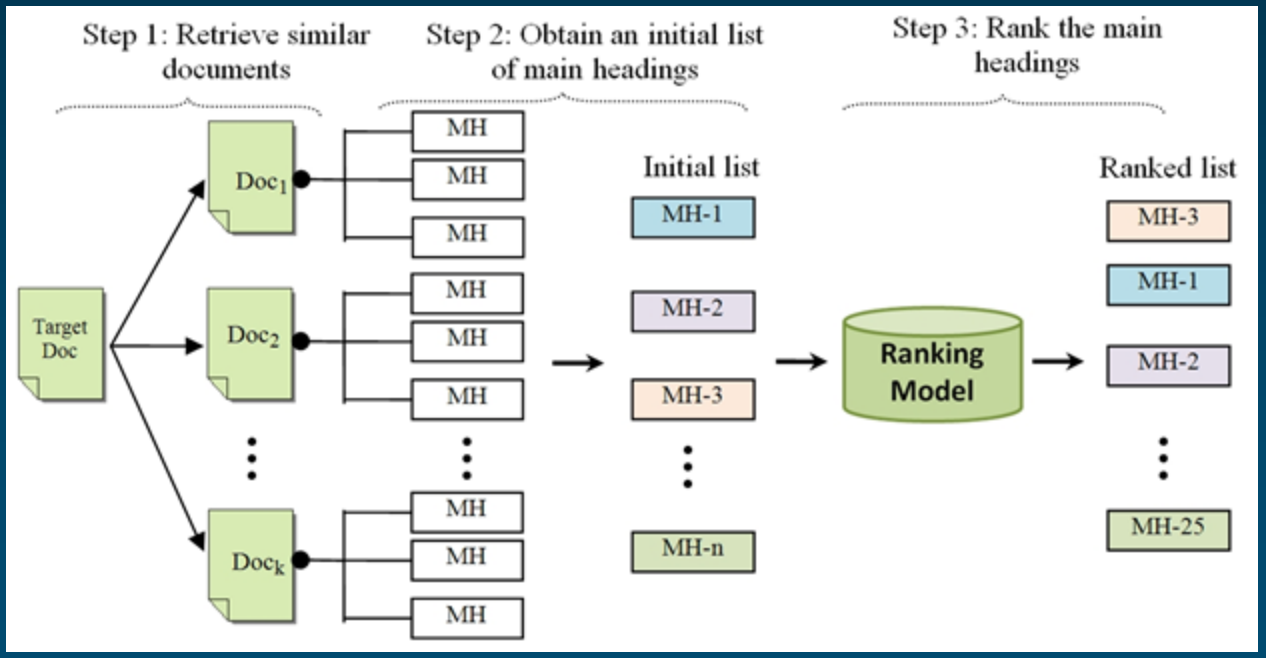
Learning to Assign MeSH Terms

Recommending MeSH terms for annotating biomedical articles
J Am Med Inform Assoc. 2011 Sep-Oct; 18(5): 660–667.

JAMIA Journal of the American Medical Informatics Association

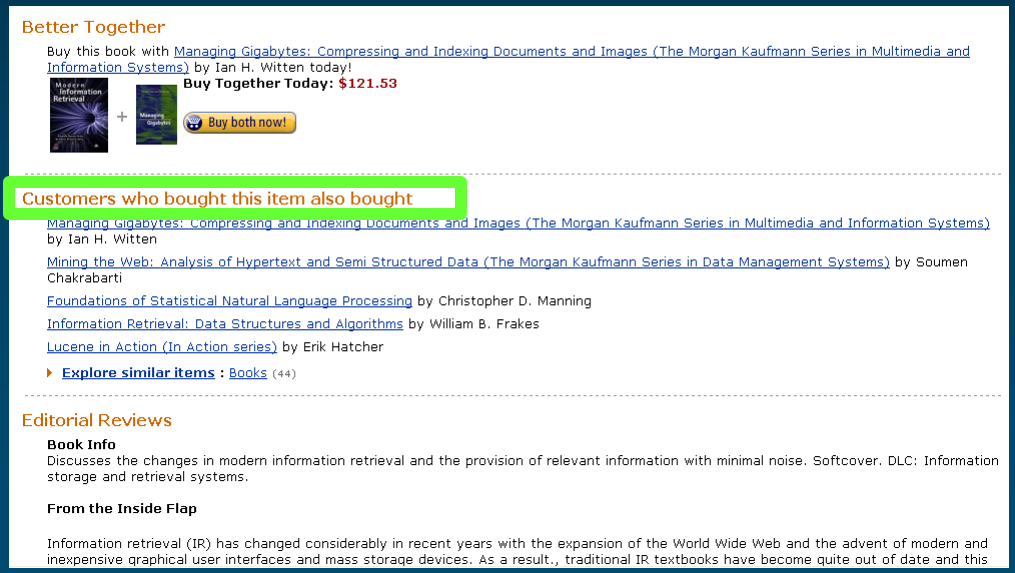
Recommending MeSH terms for annotating biomedical articles

Minlie Huang, Aurélie Névéol, and Zhiyong Lu



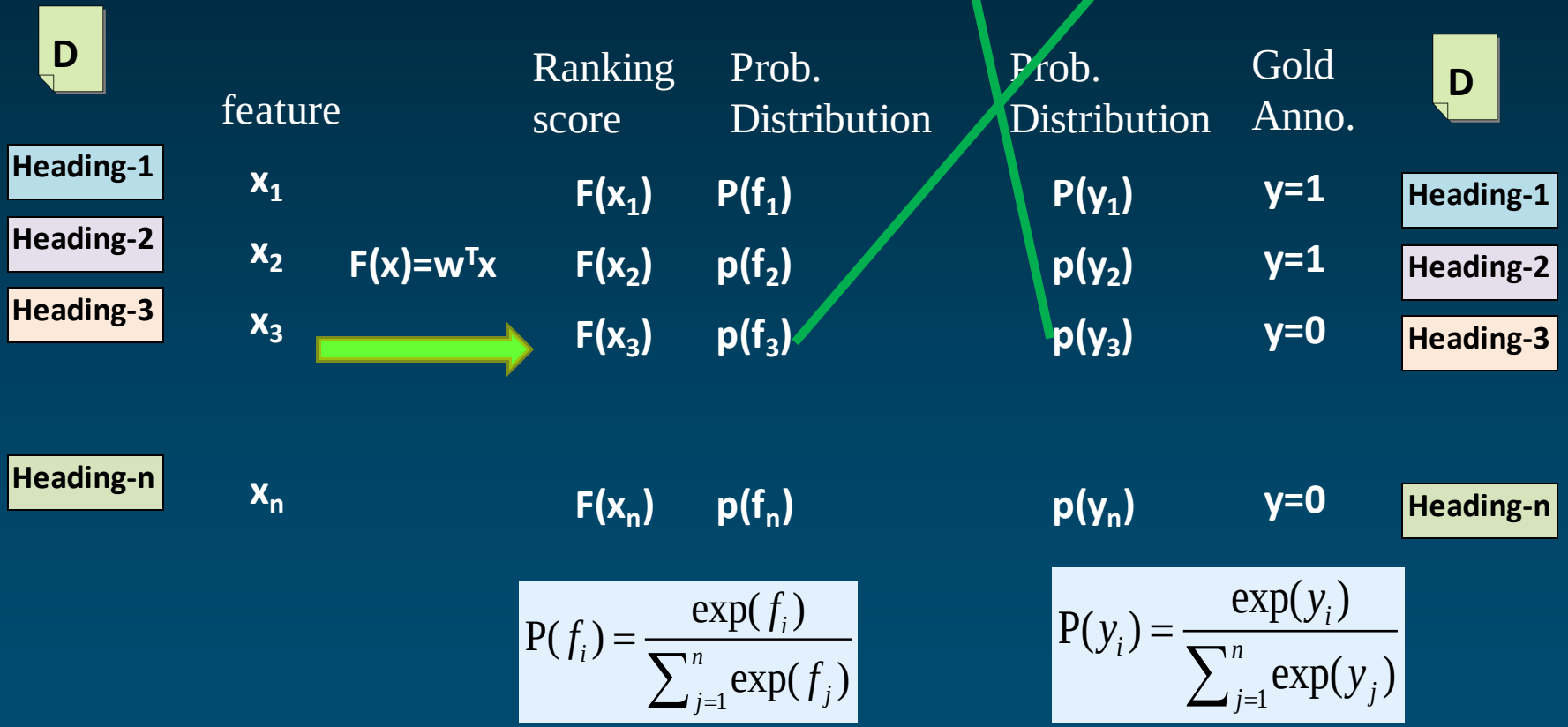
K-Nearest Neighbors

- Documents similar in content would share similar MeSH term annotations
  - Over 85% of the gold-standard MeSH annotations for a target document are present in its nearest 20 neighbors (Huang et al., 2011)



The Central Idea

$$L(Y,F) = -\sum_{j=1}^{n} \mathrm{P}(y_j) * \log \mathrm{P}(f_j)$$



$$\mathrm{P}(f_i) = \frac{\exp(f_i)}{\sum_{j=1}^{n} \exp(f_j)}$$

$$\mathrm{P}(y_i) = \frac{\exp(y_i)}{\sum_{j=1}^{n} \exp(y_j)}$$

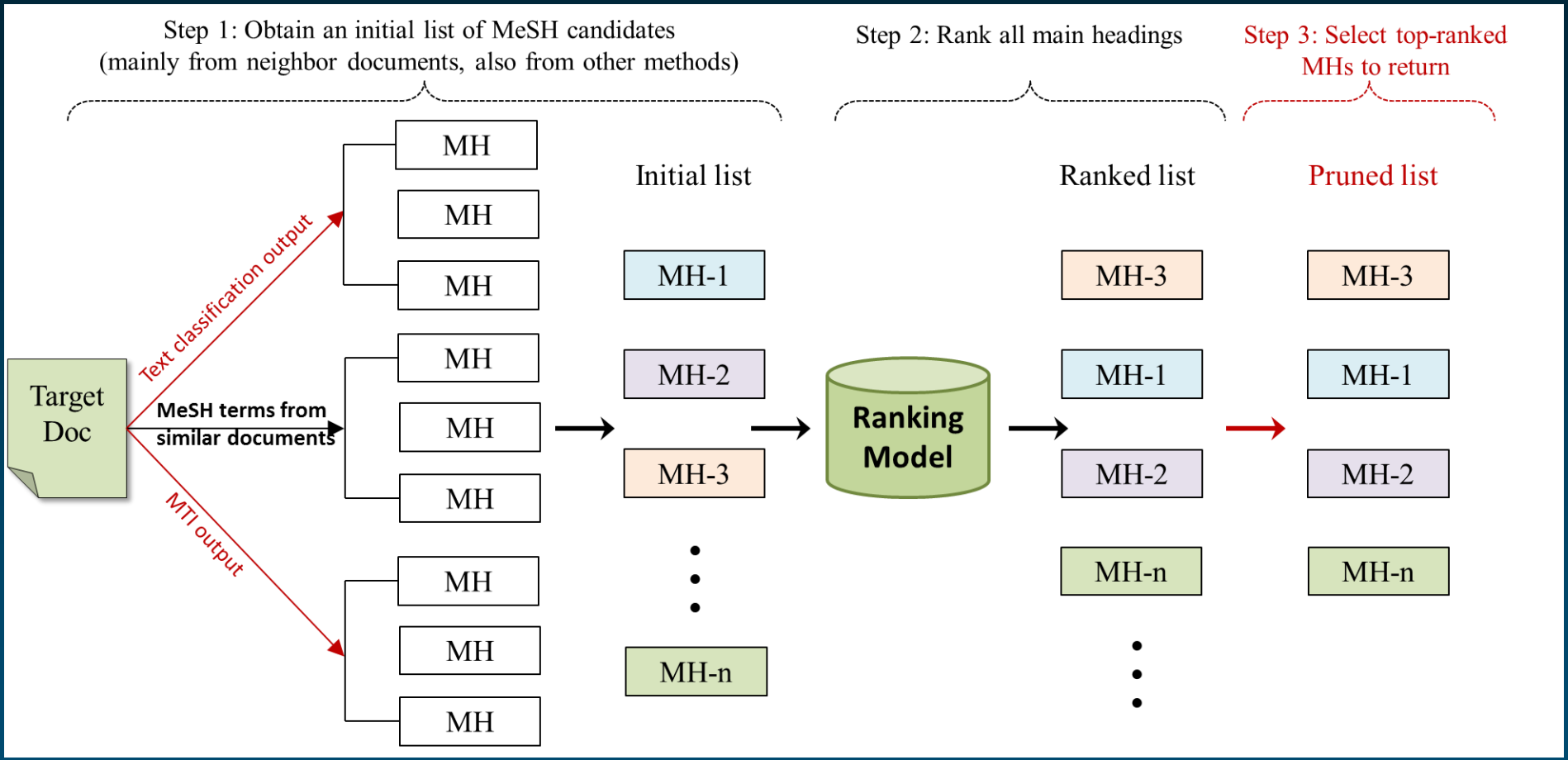
Features

- Word Unigram/Bigram Overlap Features
- Synonym Features

- Translation Probability Features
- Query-likelihood Features

- Neighborhood Features

Our 2014 System for BioASQ Task 2A



What's new in 2014

- Incorporating results from other methods
  - Binary text classification
  - NLM's MTI (Mork et al., 2013)
- System optimization
  - New L2R algorithm (LAMDA-MART)
  - MeSH 2014
  - More training data
  - List-pruning
  - Post-processing
    - E.g., Refining Age Check Tags

The NLM Medical Text Indexer System for Indexing Biomedical Literature. J.G. Mork, A. Jimeno Yepes, A.R. Aronson. 2013

BioASQ Task 2b

- Biomedical Semantic QA (involves IR, QA, summarization)
- Phase A: questions from the benchmark datasets
  - Return relevant documents, snippets, concepts, and RDF triples
- Phase B: questions and gold (correct) lists from the benchmark datasets
  - Return "exact" and "ideal" answers

Task 2b – Phase A

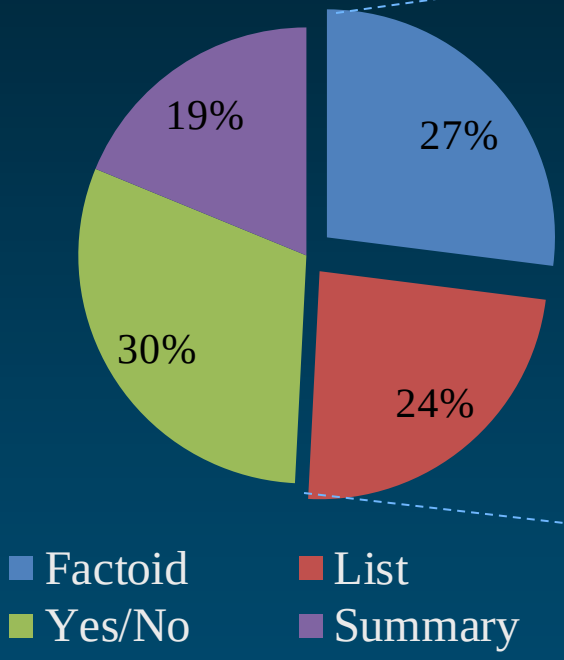
- For each natural language question, required to return:
  - Relevant documents
    - By using sort-by-date/sort-by-relevance of PubMed
  - Relevant snippets in documents
    - Compute cosine similarity score, $\cos(q,s)$, between the question (q) and each sentence (s) in a retrieved article
    - $\cos(q,s) = \frac{q \cdot s}{\|q\|\|s\|} = \frac{\sum_{t \in q \cap s} q_t \cdot s_t}{\sqrt{\sum {q_t}^2}\sqrt{\sum {s_t}^2}}$
  - Relevant concepts
    - Dictionary look-up for extracting disease, chemical and GO terms
    - GenNorm (Wei et al., 2011) for identifying gene/protein
    - MetaMap (Aronson et al., 2001) for extracting MeSH concepts
  - RDF Triples
    - Return relevant gene/protein concepts only

Wei, Kao (2011), Cross-species gene normalization by species inference. BMC Bioinformatics.
Aronson AR (2001), Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, AMIA

Task 2b – Phase B

Question type	Return answers	Example questions
Yes/No (exact/ideal answer)	Yes/No	Is there a relationship between junctin and ryanodine receptors?
Factoid (exact/ideal answer)	Each participating system will have to return a list of up to 5 entities.	Which drug is benserazide usually co-administered with?
List (exact/ideal answer)	Each participating system will have to return a list of entities	Which genes are affected in ROMANO-WARD syndrome?
Summary (ideal answer only)	Summary	What is the genetic basis of progeria?

Question/Answer Distribution



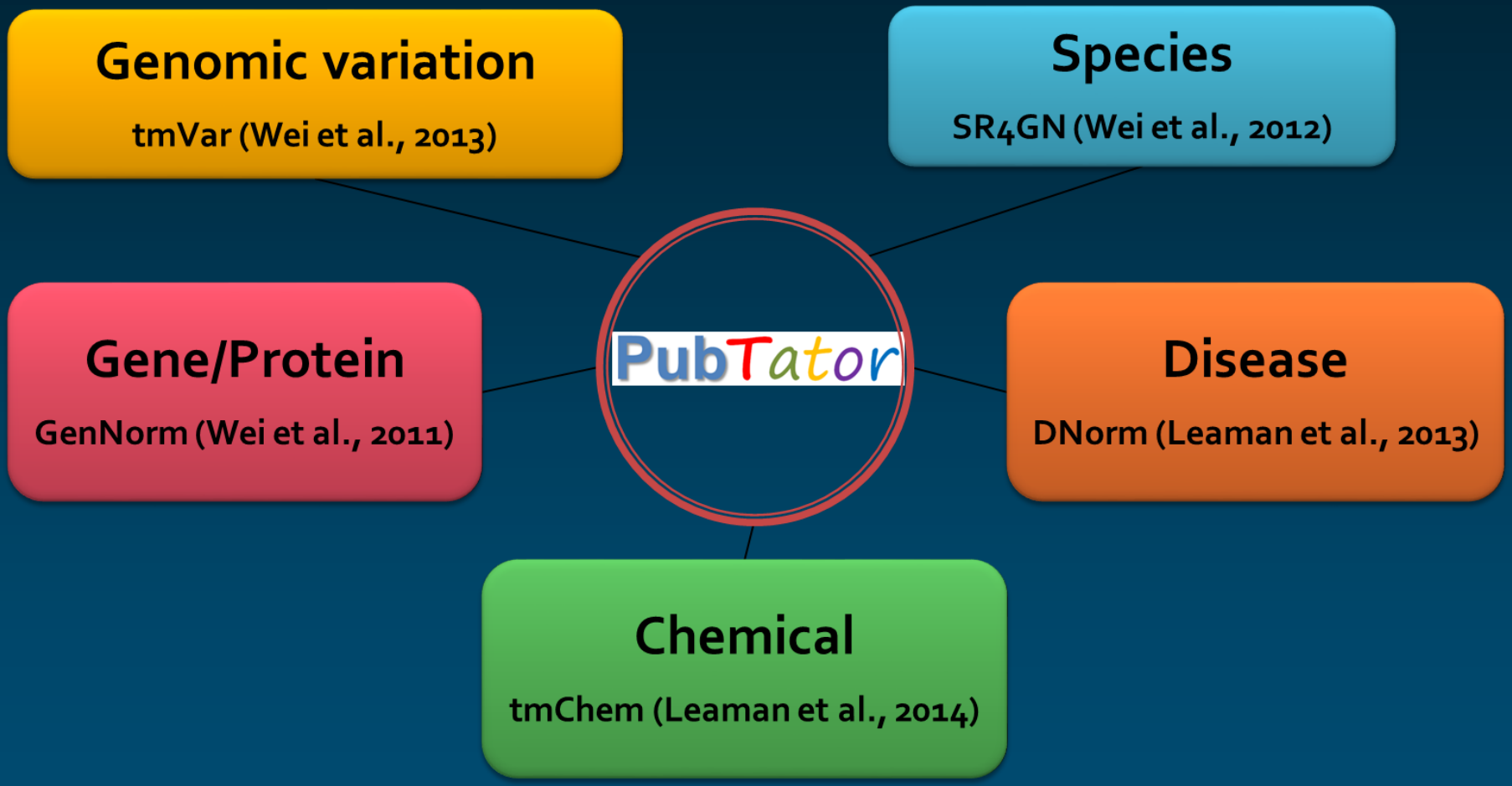
Entities		Distribution
Bio-Concepts	Gene	21.65%
	Disorder	11.42%
	mutation	3.94%
	Chemical	2.76%
	Species	0.39%
Multi-choices		1.57%
Numbers/rates		2.76%
Others (eg. motif, population)		55.51%

Factoid & List questions

Answer type	Example questions	Candidate answers
1) Bio-concepts	Which **gene** is involved in CADASIL? Which **drugs** affect insulin resistance in obesity? Which **disease** is caused by mutations in Calsequestrin 2 (CASQ2) gene? Which gene **mutations** are responsible for isolated Non-compaction cardiomyopathy? Which **virus** is Cidofovir (Vistide) indicated for?	Using PubTator* to identify bio-concepts
2) Numbers	How **many** genes does the human hoxD cluster contain? What is the **incidence** of Edwards's syndrome in the European population?	All numbers in relevant snippets
3) Multi-Choices	Is the transcriptional regulator BACH1 an **activator** or a **repressor**?	The choices in the question

*Wei CH et. al., PubTator: a Web-based text mining tool for assisting Biocuration, *Nucleic acids research*, 2013, 41 (W1): W518-W522. doi: 10.1093/nar/gkt44

PubTator is powered by text mining tools



Task 2b – Phase B: Exact Answer

- Yes/No
  - Yes (strong performance on training data)
- Factoid & List
  - Detect candidates based on three types
    - 1) Bio-concepts, 2) Numbers, 3) Multi-choice
  - Calculate the cosine similarity score of candidates (c) against snippets (s)
    - $\cos(c, s) = \frac{c \cdot s}{\|c\|\|s\|} = \frac{\sum_{t \in c \cap s} c_t \cdot s_t}{\sqrt{\sum c_t^2}\sqrt{\sum s_t^2}}$
  - Return candidates with highest scores
- Summary
  - N/A

Task 2b – Phase B: Ideal Answer

- All types of questions
  - Calculate the cosine similarity score of ideal answer candidate snippets (s) against question(q)
    - $\cos(s,q) = \frac{s \cdot q}{\|s\|\|q\|} = \frac{\sum_{t \in s \cap q} s_t \cdot q_t}{\sqrt{\sum {s_t}^2}\sqrt{\sum {q_t}^2}}$
  - Return snippets with highest scores

Official results: Task 2a

	MiF	MiP	MiR	LCA-F	LCA-P	LCA-R
NCBI(L2R-n2)	0.6076	0.6203	**0.5954**	**0.5128**	0.5338	**0.5239**
Default MTI	0.5669	0.5935	0.5425	0.4855	0.5254	0.4801
MTI First Line	0.5550	0.6270	0.4978	0.4711	0.5448	0.4414
BioASQ Baseline	0.2668	0.2414	0.2983	0.3120	0.3219	0.3301

* Test dataset: Batch 3, Week 2, size: 5,883 (the number of indexed articles by Sep. 8th)
* Training set: 5,000 articles selected from the BioASQ 2013 test sets
* Our best results among all submissions are highlighted in bold.

Official results: Task 2b Phase A

	Mean precision	Recall	F-Measure	MAP	GMAP
Documents	0.2124	0.1450	0.1384	0.0903	0.0005
Concepts	**0.4572**	0.391	**0.3848**	0.297	0.0634
RDF triples	0.0455	0.001	0.0021	0.001	0.0000
Snippets	0.0655	0.038	0.0409	0.024	0.0001

* Test dataset: Batch 5

*Our best results among all submissions are highlighted in bold.

Official results: Task 2b Phase B

- "exact" answers

Batch	Yes/No	Factoid			List		
	Accuracy	StrictAcc.	Lenient Acc.	MRR	Mean precision	Recall	F-Measure
Batch1	**0.9375**	0.1852	0.1852	0.1852	0.0618	0.0929	0.0723
Batch2	0.8214	–	–	–	0.1596	0.2057	0.1618
Batch3	0.8333	**0.0417**	**0.1250**	**0.0833**	0.1195	0.1780	0.1373
Batch4	0.8750	0.0938	0.1250	0.1042	–	–	–
Batch5	**1.0000**	**0.1379**	**0.1724**	**0.1466**	–	–	–

*Our best results among all submissions are highlighted in bold.

Official results: Task 2b Phase B

- "ideal" answers

Batch	Automatic scores		Manual scores			
	Rouge-2	Rouge-SU4	Readability	Recall	Precision	Repetition
Batch1	0.146	**0.1476**	**4.18**	3.30	3.85	**4.66**
Batch2	**0.1992**	**0.2038**	**4.28**	3.56	**4.16**	**4.68**
Batch3	**0.1592**	**0.1563**	**4.18**	3.33	3.73	**4.57**
Batch4	0.1648	0.1744	**4.00**	3.67	3.86	**4.45**
Batch5	**0.1671**	**0.1775**	**4.06**	3.81	4.07	**4.45**

*Our best results among all submissions are highlighted in bold.

Discussion

- Task 2a
  - Improved performance
  - Highest recall after including other sources
  - Performance ceilings
    - Check Tags from full-text
- Task 2b
  - Best in the “exact” and “ideal” answers
    - to the Factoid-type questions
  - Used results of our previously developed NER tools
    - better than relying on the gold-standard concepts from Phase A

Conclusions

- BioASQ 2014 challenge
  - among the winner teams for both tasks
- A general and robust framework
  - allows systematic integration of results from other methods for improved performance
- MeSH prediction
  - in practical applications
- Question answering
  - automated entity recognition tools

Acknowledgements

- BioASQ task organizers
  - MTI team: James Mork, Alan Aronson
- This research is supported by the NIH Intramural Research Program, National Library of Medicine
  - Dr. Ritu Khare
  - Dr. Robert Leaman

NCBI Text Mining Tools (tmTools)

http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools

Zhiyong Lu (zhiyong.lu@nih.gov)

Yuqing Mao (yuqing.mao@nih.gov)

NCBI, NLM, NIH

Bethesda, Maryland - USA