BioASQ Workshop, 27 September 2013, Valencia, Spain

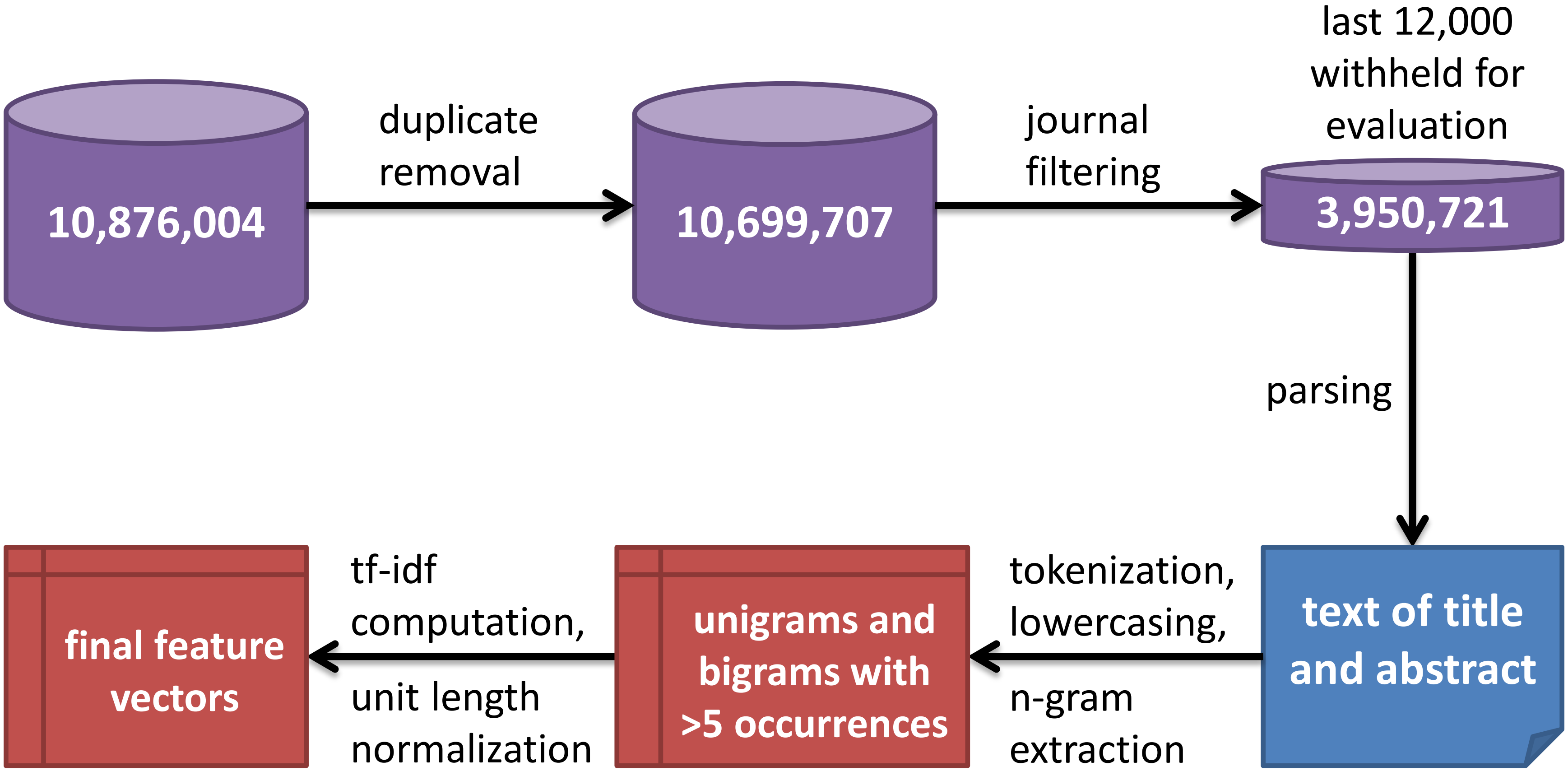
Large-Scale Semantic Indexing of Biomedical Publications at BioASQ

Grigorios Tsoumakas[1], Manos Laliotis[2a], Nikos Markantonatos[2b], Ioannis Vlahavas[1]

[1] Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki 54124, Greece

[2] Atypon,
[2a] Santa Clara, CA 95054, USA
[2b] Athens, 15341, Greece

Preprocessing



10,876,004

duplicate removal

10,699,707

journal filtering

last 12,000 withheld for evaluation

3,950,721

parsing

text of title and abstract

tokenization, lowercasing, n-gram extraction

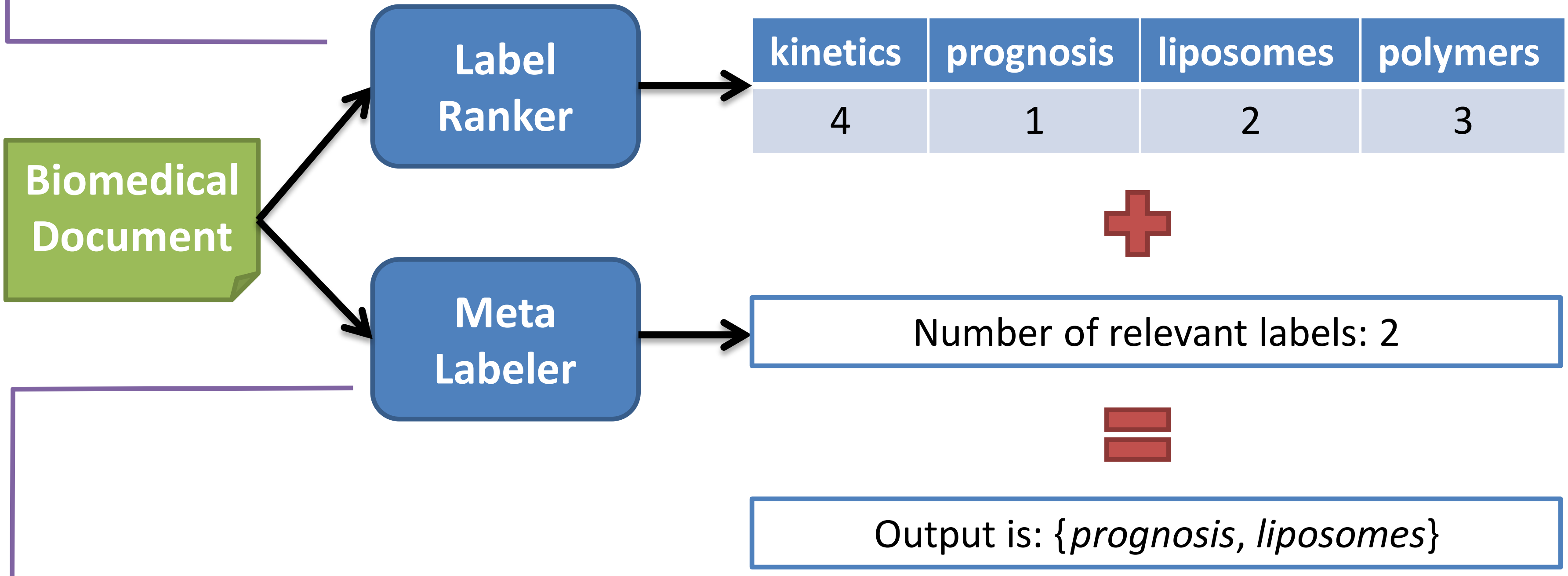
unigrams and bigrams with >5 occurrences

tf-idf computation, unit length normalization

final feature vectors

Meta-Labeler (*Tang et al., WWW09*)

- Any multi-label learning algorithm that can output a ranking of the labels
- We used a linear SVM per label and considered their unthresholded output



kinetics	prognosis	liposomes	polymers
4	1	2	3

Number of relevant labels: 2

Output is: {*prognosis*, *liposomes*}

- Regression or (ordinal) classification using original features or label scores/ranks
- We used linear SVM regression based on the original features

Tried But Failed

- Thresholding
  - SCut
- Class imbalance counterfeiting
  - Majority under-sampling, asymmetric bagging
- Representations
  - Plain unigrams/bigrams, addition of trigrams
  - Bi-normal separation (BNS) scaling
- Hierarchy exploitation
  - Top-down approach

Our 4 Particular Systems

- Systems 1 and 3
  – Default meta-labelers as presented two slides ago
- System 2
  – Binary SVMs for some labels and meta-labeler for others, cyclically optimized on evaluation set
- System 4
  – Majority voting of 3 default meta-labelers

System	Publications	Unigrams	Bigrams	Labels
1	800,000	215,133	1,908,088	25,625
2,3	700,000	197,590	1,720,818	25,509
4	500,000 x 3	138,196±30	1,097,465±188	25,214±1.4

Time and Space

- Hardware
  - 4 10-core processors at 2.26 GHz, 1 Tb RAM and 2.4 Tb storage (6 x 600 Gb SAS 10k disks in RAID 5)
- Parallel learning/use of binary SVMs
  - Using 40 threads of a 4 10-core processor system at 2.26 GHz, training took a couple of days, while prediction took a couple of hours
- Serialization
  - Required to respond within the 16h limit
  - Storing the models of system 1 required 406 Gb

Results

- System 1 topped F-measure
  - Best in Micro and Lowest Common Ancestor (LCA) F-measure from its introduction in c2w4 till c3w5
  - Micro F-measure: ~0.57, LCA F-measure: ~0.48
- System 2 generally worse than System 3
  - Need to reconsider selection of models per label
  - Binary SVMs better especially in frequent labels
- System 4 topped precision
  - 0.83 (example-based), 0.82/0.76 (micro/macro)
  - 0.91 (hierarchical), 0.56 (LCA)

Open Issues

- Considering the temporal dimension of data
  - Frequency of labels varies over time
  - Can techniques for handling concept drift help?
- Considering journal information
  - E.g. using journal title as text
- Handling title and abstract text separately
  - Zoning, different vocabularies

BioASQ Workshop, 27 September 2013, Valencia, Spain

Large-Scale Semantic Indexing of Biomedical Publications at BioASQ

Grigorios Tsoumakas[1], Manos Laliotis[2a], Nikos Markantonatos[2b], Ioannis Vlahavas[1]

[1] Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki 54124, Greece

[2] Atypon,
[2a] Santa Clara, CA 95054, USA
[2b] Athens, 15341, Greece