An Ensemble Approach for the BioASQ challenge 2014

Yannis Papanikolaou, Dimitrios Dimitriadis, Grigorios Tsoumakas,
Manos Laliotis , Nikos Markantonatos and Ioannis Vlahavas

BioASQ Task 2a

Challenges:

- Multilabel problem
  - dependencies among labels
- Scalability
  - tens of thousands of labels, millions of documents
- Time limitation for required results
  - <24 hours
- Concept drift
  - concepts change over time

Our Approach

- Employ state-of-the-art methods both discriminative (SVM, MetaLabeler) and probabilistic (Labeled LDA)
- Implement a new multilabel ensemble method
- Keep it simple, fast and scalable (w.r.t the previously mentioned challenges)

Component Models

- SVM
- MetaLabeler
- Labeled LDA

SVM (1/2)

- Binary relevance (BR) approach, each label is learnt/predicted separately not taking into account depedencies
- Simple
- Parallellizable
- Extremely scalable
- LibLinear implementation[1]

1. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. J. Mach. Learn. Res. 9 (June 2008) 1871–1874

SVM (2/2)

- Vanilla SVM : no parameter tuning at all
- “Tuned” SVM : handle class imbalance by penalizing more heavily false negative(fn) errors than false positive (fp) errors[1]
- Feature selection and BNS scaling also tried but proved unsuccessful

1. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. J. Mach. Learn. Res. 5 (2004) 361–397

MetaLabeler[1]

Idea:

doc #1

ℓ_3

ℓ_2

$n = 3$ ℓ_6

ℓ_1

doc #2

ℓ_5

ℓ_3 $n = 1$

ℓ_2

ℓ_4

ℓ_i :score of label i, n:threshold

- When k-fold cross-validation is difficult (large data) we can train a simple regression model to determine number of labels per instance

1. Tang, L., Rajan, S., Narayanan, V.K.: Large scale multi-label classification via metalabeler. In: WWW '09: Proceedings of the 18th international conference on World wide web, New York, NY, USA, ACM (2009) 211–220

Labeled LDA[1, 2]

- Probabilistic background
- Supervised approach of LDA
- Idea: learn the φ(l, w) distributions (labels-words) during training and compute the θ(l, d) distributions (labels – documents) during inference.

1. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1. EMNLP '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 248–256
2. Rubin, T.N., Chambers, A., Smyth, P., Steyvers, M.: Statistical topic models for multi-label document classification. Mach. Learn. 88(1-2) (July 2012) 157–208

Performance of Component Models

Model	miF
Vanilla SVM	0.56192
Tuned SVM	0.58330
MetaLabeler+Vanilla SVM	**0.59461**
Labeled LDA	0.38321

Results for the models trained on 1.5 million documents of the BioAsq corpus and tested on 35k annotated documents from the competition batches

MULtilabel Ensemble (MULE)

1..ℓ labels,

A: baseline model

B_i: other models

- Classifier selection scheme
- Idea:

  a) for every ℓ, substitute model's A prediction ($tp_{A\ell}$, $fp_{A\ell}$, $tn_{A\ell}$, $fn_{A\ell}$) with model's B_i prediction ($tp_{B\ell}$, $fp_{B\ell}$, $tn_{B\ell}$, $fn_{B\ell}$) and check if this improves total performance (miF)

  b) use a significance test to validate the selection

Example 1/4

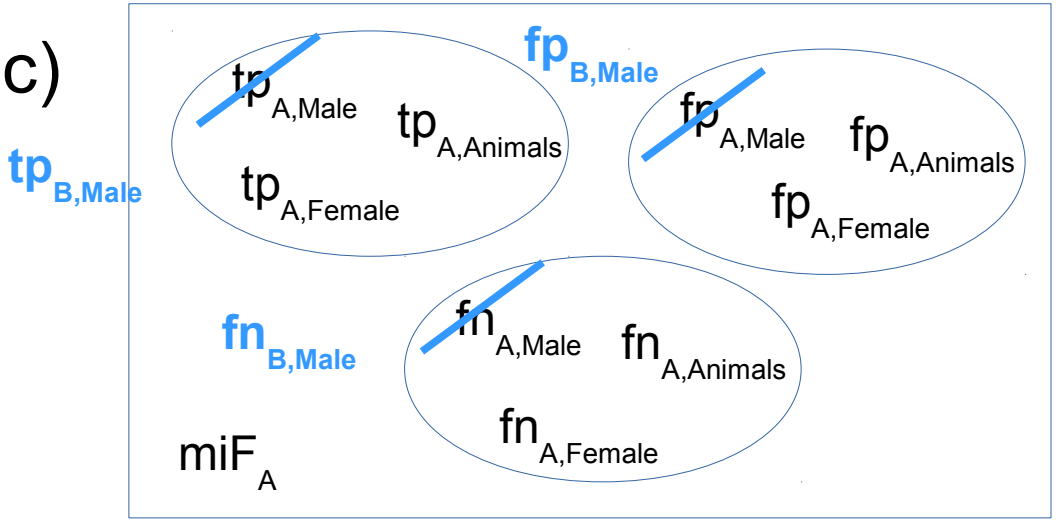
- Models: A, B_1, B_2 (for simplicity A is the best performing one)
- Labels: *Male, Female, Animals*

Example 2/4

a) Train all models on a `training_set`

b) Compute miF_A on a `validation_set`

c)

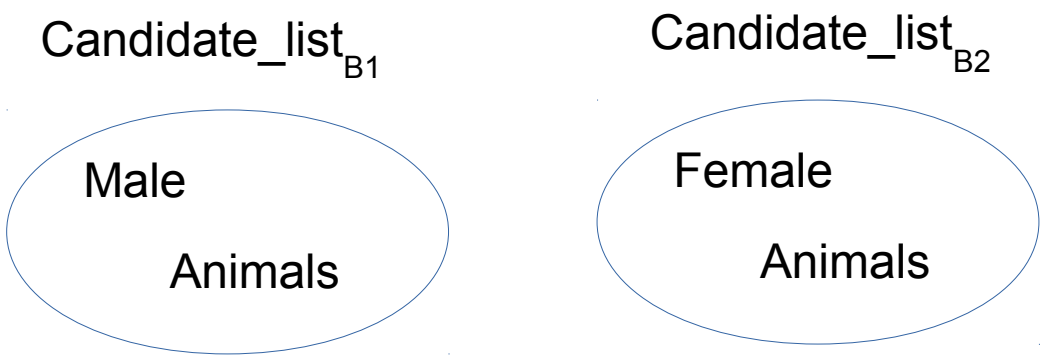


d) If $miF_{A,\ B1 \sim male} > miF_A$ then add *Male* to $candidate\_list_{B1}$

Example 3/4

e) Repeat (c) - (d) for all labels and all models

f)



g) McNemar tests: A-B_1 for “Male”, A-B_2 for “Female”, A-B_1-B_2 for “Animals” (A-B_1 & A-B_2)

Example 4/4

h) Suppose A-B_1 difference for “Male” is s.s. , A-B_2 for “Female” is not and both B_1, B_2 are s.s. better than A; Then predict:

- “Male” → B_1
- “Female” → A
- “Animals” → B_1 or B_2 (whichever performs better - we don't care about s.s.)

Notes

- Reliable even for small v.datasets, but perhaps a bit conservative in this case
- Ommitting the statistical test leads to non reliable results
- Selecting classifiers with F instead of miF brings negative results even when testing on the v.dataset[1]

1. Jimeno-Yepes, A., Mork, J.G., Demner-Fushman, D., Aronson, A.R.: A one-size-fits-all indexing method does not exist: Automatic selection based on meta-learning. JCSE 6(2) (2012) 151–160

Performance of Systems

Systems	miF
Hippocrates/Asclepios (MetaLabeler)	0.60921
Sisyphus(MetaLabeler+Tuned SVMs)	0.61923
Galen (MetaLabeler+LLDA)	0.60949
Panacea (MetaLabeler+Tuned SVMs+LLDA)	0.61968

Results are shown for 12.3k documents, having used 35k documents for validation and 1.5m for training. The ensemble systems perform better than the baseline (Hippocrates), even if the validation data set is relatively small.

Conclusions & Future Work

- The new multilabel ensemble method we proposed proved successful both in our experiments and the BioAsq challenge (1st place on the first batch, 3rd on the two others)
- Possible future work could include:
  a) Use of other thresholding approaches[1, 2]
  b) Improvements over the labeled LDA algorithm (parameter tuning, parallelization, etc)

1.Tahir, M.A., Kittler, J., Bouridane, A.: Multilabel classification using heterogeneous ensemble of multi-label classifiers. Pattern Recogn. Lett. 33(5) (2012) 513–523
2.Nam, J., Kim, J., Gurevych, I., Furnkranz, J.: Large-scale multi-label text classification -revisiting neural networks. CoRR abs/1312.5419 (2013)

BioASQ Task 2B – Phase B

- Newcomers replicating last year's work[1]
- Ensemble of 5 scores of candidate answers
  - (Weighted) Prominence, Specificity
  - TypeCoercionLAT, TypeCoercionQuestion

Scoring	SAcc	LAcc	MRR
Prominence (P)	9%	31%	16%
WeightedProminence (WP)	23%	31%	25%
Specificity (S)	4%	23%	11%
P + WP + S	31%	43%	35%
P + WP + S + TypeCoercionLAT (TCLAT)	26%	40%	31%
P + WP + S + TCLAT × 0.5	29%	45%	35%
P + WP + S + TypeCoercionQuestion (TCQ)	24%	45%	33%
P + WP + S + TCQ × 0.5	29%	48%	36%
P + WP + S + TCQ × 0.5 + TCLAT	24%	43%	32%
P + WP + S + TCQ + TCLAT × 0.5	24%	48%	35%

1. Weissenborn, D., Tsatsaronis, G., Schroeder, M.: Answering factoid questions in the biomedical domain. In Ngomo, A.C.N., Paliouras, G., eds.: BioASQ@CLEF.Volume 1094 of CEUR Workshop Proceedings., CEUR-WS.org (2013)

An Ensemble Approach for the BioASQ challenge 2014

Questions