An Ensemble Approach for the BioASQ challenge 2014

Yannis Papanikolaou, Dimitrios Dimitriadis, Grigoris Tsoumakas, Manos Laliotis, Nikos Markantonatos and Ioannis Vlahavas
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 dependencies
- Simple
- Parallelizable
- Extremely scalable
- LibLinear implementation

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

• Feature selection and BNS scaling also tried but proved unsuccessful

MetaLabeler

Idea:

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

\[ \ell_3 \]
\[ \ell_2 \]
\[ \ell_6 \]
\[ \ell_1 \]

\[ \ell_5 \]
\[ \ell_3 \]
\[ \ell_2 \]
\[ \ell_4 \]

\( n = 3 \)

\( n = 1 \)

\( \ell_i \) : score of label i, n : threshold

Labeled LDA$^1, 2$

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

### Performance of Component Models

<table>
<thead>
<tr>
<th>Model</th>
<th>miF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla SVM</td>
<td>0.56192</td>
</tr>
<tr>
<td>Tuned SVM</td>
<td>0.58330</td>
</tr>
<tr>
<td>MetaLabeler+Vanilla SVM</td>
<td>0.59461</td>
</tr>
<tr>
<td>Labeled LDA</td>
<td>0.38321</td>
</tr>
</tbody>
</table>

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
Bᵢ: other models

• Classifier selection scheme
• Idea:
  a) for every ℓ, substitute model's A prediction \((tp_{Aℓ}, fp_{Aℓ}, tn_{Aℓ}, fn_{Aℓ})\) with model's Bᵢ prediction \((tp_{Bℓ}, fp_{Bℓ}, tn_{Bℓ}, fn_{Bℓ})\) 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 $\text{miF}_A$ on a validation_set
c) 

d) If $\text{miF}_{A, B_1 \sim \text{male}} > \text{miF}_A$ then add $\text{Male}$ to candidate_list$_{B_1}$
Example 3/4

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

f) Candidate_list_{B_1}
   \begin{itemize}
   \item Male
   \item Animals
   \end{itemize}

   Candidate_list_{B_2}
   \begin{itemize}
   \item Female
   \item Animals
   \end{itemize}


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)
h) Suppose A-B₁ difference for “Male” is s.s., A-B₂ for “Female” is not and both B₁, B₂ are s.s. better than A; Then predict:
- “Male” $\rightarrow B₁$
- “Female” $\rightarrow A$
- “Animals” $\rightarrow B₁$ or $B₂$ (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
• Omitting 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\(^1\)

### Performance of Systems

<table>
<thead>
<tr>
<th>Systems</th>
<th>miF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hippocrates/Asclepios (MetaLabeler)</td>
<td>0.60921</td>
</tr>
<tr>
<td>Sisyphus(MetaLabeler+Tuned SVMs)</td>
<td>0.61923</td>
</tr>
<tr>
<td>Galen (MetaLabeler+LLDA)</td>
<td>0.60949</td>
</tr>
<tr>
<td>Panacea (MetaLabeler+Tuned SVMs+LLDA)</td>
<td>0.61968</td>
</tr>
</tbody>
</table>

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\textsuperscript{1}, \textsuperscript{2}
  b) Improvements over the labeled LDA algorithm (parameter tuning, parallelization, etc)

BioASQ Task 2B – Phase B

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

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

An Ensemble Approach for the BioASQ challenge 2014

Questions