PubMeSH: Extreme multi-label classification of biomedical research

CS224N
Kevin Thomas¹, Rohan Paul¹, Mia Kanzawa¹
1. Department of Biomedical Data Science, Stanford University, Stanford, CA.

Problem

- PubMed/MEDLINE is central hub for biomedical and life sciences journal articles and critical for dissemination of scientific knowledge
  - Contains >29 million articles
  - 3.3 billion searches in 2017
- Searches rely on MeSH labels which are manually assigned to each article
  - Estimated >$7 million in labor costs/year
- Automatic labelling could significantly reduce costs
- Previous state-of-the-art: LSTM with attention
  - MiF of 0.6880 → requires human review
  - Limit # of possible labels with masking approach (misses up to 5% of labels)
  - Naïve representation of MeSH labels

Data/Task

Goal: assign a set of MeSH terms to each abstract
- multi-label, multi-class classification

- >14 million abstracts
- >29,000 MeSH terms
- ~10-15 MeSH terms/article

Approach

- Frequency-based Masking
- Pre-trained Embedding

Results

<table>
<thead>
<tr>
<th>Model Name</th>
<th>MiF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common</td>
<td>0.670</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.384</td>
</tr>
<tr>
<td>Rare</td>
<td>0.229</td>
</tr>
<tr>
<td>Super Rare</td>
<td>0.171</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.484</td>
</tr>
</tbody>
</table>

Analysis

- Produces viable labels absent from ground truth
  - [Curriculum, Education, Internship and Residency, Students]
    - [Clinical Competence, Education]
  - [Oxygen, Exercise]
    - [Oxygen Consumption, Physical Endurance, Heart Rate]
  - [Fracture Fixation, Osteoporosis]
    - [Lumbar Vertebrae, Pain Measurement]

Misses important labels

- [Brazil, Food, Fruit]
  - [Brazil, Fruit, Meat, Vegetables, Energy Intake, Eating, Feeding Behavior, Obesity, Body Mass Index]

Future directions:
  - Ablation studies for linear layers
  - Freezing embedding layers
  - Transfer learning
  - Hyperparameter tuning

Conclusion

- Proposed novel methods for automated MeSH indexing do not yet surpass state-of-the-art
  - Pre-trained MeSH embeddings
  - Separate training for prediction of MeSH labels in different frequency categories

References