The mycoSORT System

mycoSORT utilizes supervised machine learning to perform automatic text classification of PubMed abstracts. The system goal is to support the triage of candidate articles for the mycoCLAP database [1]. Over 100 classification models were evaluated to identify the best performance under different settings. The purpose is to be able to handle the triage task under an imbalanced class distribution.

mycoSORT Pipeline

Training phase
1. Input curated abstracts
2. Extraction of relevant features
3. Classification model building

Testing phase
1. Input unlabeled abstracts
2. Extraction of relevant features
3. Classification model application
4. Evaluation of system predictions

Feature Extraction

- Pre-processing: stop-words, ASCII characters, markup tags
- Content: abstract, paper title and Enzyme Comission (EC) numbers
- Annotation spans: entity and sentence
- Filtering criteria: length > 3 and occurrence > 2

Feature Vector

| substrate | specificity | three | enzyme | ligninase | isozymes | whiterot | fungus | Trametes versicolor | investigated | I1499 | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Bioentities of the entity span
[ligninase, Enzyme], [Trametes versicolor, fungus]

Bioentities of the sentence span
[substrate, substratespecificity], [specificity, substratespecificity], [three, substratespecificity], [ligninase, substratespecificity], [ozymes, substratespecificity], [whiterot, substratespecificity], [fungus, substratespecificity], [Trametes versicolor, substratespecificity], [investigated, substratespecificity]

Dataset

- Over 7,580 manually curated PubMed abstracts
- 749 relevant (POS) and 6,834 not relevant (NEG)
- Evaluation of several class distributions
- Training: random undersampling of majority class
- Testing: real class distribution (10%POS, 90%NEG)
**Experimental Settings**

**Classification algorithms**
- Naive Bayes (NB)
- Support Vector Machine (SVM)
- Logistic Model Trees (LMT)

**Feature Settings**
- #1: F1
- #2: F1 + F4
- #3: F5
- #4: F1 + F2 + F3 + F4

**Undersampling Factors**
- 0% USF: 90%NEG, 10%POS
- USFs gradually increased by 5%
- 40% USF: 50%NEG, 50%POS

**mycoSORT results**

![Graph showing F-measure and F-2 score vs. USF (%)]

- Baseline: #3 (bag-of-words)
- #3 feature space: from 7,622 to 20,729
- #4 feature space: from 3,338 to 8,931
- #4 outperforms #3 by using domain annotations

**Discussion**

**Future Work**

- Study the statistical relevance of features
- Use feature selection to refine existing models
- Evaluate system on different biomedical datasets

**References**