DDSI Internship Experience

7/22/20
Will Hibbard
Background/ Overview

- The UMLS metathesaurus is a database of medical vocabularies and standards

- Goal: to semantically group the atoms in the UMLS database based on lexical and contextual information to allow for better synonymy prediction

- Why: sorting the terms into their semantic groups allows for better synonymy prediction with the atoms, but previous lexical only methods caused false synonymy between terms
  - Ex. Splint (shin) and splint (medical device)

- How: Using deep learning, we embed atoms into vectors and train a neural network on those until it can accurately predict the semantic groups of those atoms. Then, the model is tested with a dataset it’s never seen before
Learning Neural Networks

• I started by reading up on Neural Networks and going over example code with Yuqing
• The code I based my neural network off of was from a programming blog that made a neural network to classify wines by type
• Admittedly, I spent the first few weeks getting access to TOAD and Biowulf
• My network ended up being a 5-layer classification model with SAP BERT embedding
• SAP BERT was chosen because it maintained the sentence and paragraph structure of the atoms it embedded, allowing for contextual info like word placement and order to be considered
The Dataset

• Gathered with TOAD on a virtual machine
• The desired dataset could only have atoms that were in English and came from active source vocabularies
• The information in the dataset was AUI’s, the string, and TUIs
  • AUIs were used as unique identifiers
  • The string was the data we wanted
  • TUIs were used to map the atoms to their correct groups so the model could gauge its performance
• The entire dataset was 10,000,000+ atoms and was too big to run normally, so two subsets were made for code training and testing
The Subsets

• The subsets used to train the model were ~1,000,000+ atoms, and were quicker and more manageable to run

• The subsets were assembled with stratified sampling to ensure that the model had practice sorting every one of the semantic groups

• 10 % subset: This subset was assembled with 10% of each SG to give the model something small to run that was proportionally accurate to the whole set

• Balanced subset: This subset was made of 7000 atoms from each SG to see how the model performed without bias weights towards larger SGs
Experimentation

- Experiments for the classification problem were done by running the datasets with different values for Epoch, Batch Size, Learning Rate, Hidden Layers, and nodes per Hidden Layer
  - Default: Epoch = 30, Batch Size = 512, Learning Rate = 0.0007, # of hidden layers = 3, nodes per layer = 64, 128, 512, etc…
- The parameters were adjusted individually to prevent overfitting the model
- After the optimal hyperparameters were collected, they were tested together on the datasets
• The results of each run were broken down into precision, recall, F-1, average, macro average, and weighted average
  • Precision – the model’s accuracy in classifying an atom to a group
  • Recall – was the positive classification correct
  • F-1 – a score that says the weighted average of precision and recall
  • Accuracy – how the model did across all classes
  • Macro average – the unweighted average of each column
  • Weighted average – the average that accounts for the class support

• The supports don’t reflect the whole subset size as the code divides the subset into training, validation, and testing sets

<table>
<thead>
<tr>
<th>Epoch=30, Batch size=512</th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTI</td>
<td>0.19</td>
<td>0.52</td>
<td>0.27</td>
<td>400</td>
</tr>
<tr>
<td>ANAT</td>
<td>0.34</td>
<td>0.68</td>
<td>0.45</td>
<td>9363</td>
</tr>
<tr>
<td>CHEM</td>
<td>0.89</td>
<td>0.50</td>
<td>0.64</td>
<td>53736</td>
</tr>
<tr>
<td>CONC</td>
<td>0.08</td>
<td>0.50</td>
<td>0.14</td>
<td>5013</td>
</tr>
<tr>
<td>DEVI</td>
<td>0.22</td>
<td>0.75</td>
<td>0.34</td>
<td>2499</td>
</tr>
<tr>
<td>DISO</td>
<td>0.85</td>
<td>0.47</td>
<td>0.61</td>
<td>38546</td>
</tr>
<tr>
<td>GENE</td>
<td>0.43</td>
<td>0.88</td>
<td>0.58</td>
<td>8932</td>
</tr>
<tr>
<td>GEOG</td>
<td>0.04</td>
<td>0.24</td>
<td>0.06</td>
<td>424</td>
</tr>
<tr>
<td>LIVB</td>
<td>0.88</td>
<td>0.46</td>
<td>0.60</td>
<td>50503</td>
</tr>
<tr>
<td>OBJC</td>
<td>0.07</td>
<td>0.39</td>
<td>0.11</td>
<td>1124</td>
</tr>
<tr>
<td>OCCU</td>
<td>0.07</td>
<td>0.24</td>
<td>0.11</td>
<td>400</td>
</tr>
<tr>
<td>ORGA</td>
<td>0.07</td>
<td>0.29</td>
<td>0.11</td>
<td>400</td>
</tr>
<tr>
<td>PHEN</td>
<td>0.12</td>
<td>0.33</td>
<td>0.18</td>
<td>800</td>
</tr>
<tr>
<td>PHYS</td>
<td>0.73</td>
<td>0.85</td>
<td>0.79</td>
<td>12588</td>
</tr>
<tr>
<td>PROC</td>
<td>0.59</td>
<td>0.61</td>
<td>0.60</td>
<td>18699</td>
</tr>
</tbody>
</table>

accuracy                           0.54    203427
macro avg acc       0.37      0.51    0.37     203427
weighted avg acc      0.76      0.54    0.59   203427
set = 10% total subset
Time=~20 min
Results

• Some incremental improvement in accuracy across optimal hyperparameters, but the overall accuracy was not practically applicable
• I was unable to get the whole dataset running until this week, so the results depict the subsets
### Results pt. 2

**Epoch = 30, BatchSize = 512**

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTI</td>
<td>0.38</td>
<td>0.34</td>
<td>0.36</td>
<td>1400</td>
</tr>
<tr>
<td>ANAT</td>
<td>0.54</td>
<td>0.59</td>
<td>0.56</td>
<td>1400</td>
</tr>
<tr>
<td>CHEM</td>
<td>0.49</td>
<td>0.56</td>
<td>0.52</td>
<td>1400</td>
</tr>
<tr>
<td>CONC</td>
<td>0.35</td>
<td>0.31</td>
<td>0.33</td>
<td>1400</td>
</tr>
<tr>
<td>DEVI</td>
<td>0.60</td>
<td>0.58</td>
<td>0.59</td>
<td>1400</td>
</tr>
<tr>
<td>DISO</td>
<td>0.52</td>
<td>0.45</td>
<td>0.48</td>
<td>1400</td>
</tr>
<tr>
<td>GENE</td>
<td>0.72</td>
<td>0.71</td>
<td>0.71</td>
<td>1400</td>
</tr>
<tr>
<td>GEOG</td>
<td>0.34</td>
<td>0.35</td>
<td>0.35</td>
<td>1400</td>
</tr>
<tr>
<td>LIVB</td>
<td>0.62</td>
<td>0.58</td>
<td>0.60</td>
<td>1400</td>
</tr>
<tr>
<td>OBJC</td>
<td>0.36</td>
<td>0.39</td>
<td>0.37</td>
<td>1400</td>
</tr>
<tr>
<td>OCCU</td>
<td>0.38</td>
<td>0.35</td>
<td>0.37</td>
<td>1400</td>
</tr>
<tr>
<td>ORGA</td>
<td>0.32</td>
<td>0.36</td>
<td>0.34</td>
<td>1323</td>
</tr>
<tr>
<td>PHEN</td>
<td>0.36</td>
<td>0.37</td>
<td>0.37</td>
<td>1400</td>
</tr>
<tr>
<td>PHYS</td>
<td>0.74</td>
<td>0.84</td>
<td>0.79</td>
<td>1400</td>
</tr>
<tr>
<td>PROC</td>
<td>0.41</td>
<td>0.36</td>
<td>0.38</td>
<td>1400</td>
</tr>
</tbody>
</table>

**Accuracy** 0.48 20923

**Macro Avg** 0.48 20923

**Weighted Avg** 0.48 20923

*Set = balanced data set*

**Epoch = 100, BatchSize = 256**

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTI</td>
<td>0.33</td>
<td>0.37</td>
<td>0.35</td>
<td>1400</td>
</tr>
<tr>
<td>ANAT</td>
<td>0.62</td>
<td>0.54</td>
<td>0.58</td>
<td>1400</td>
</tr>
<tr>
<td>CHEM</td>
<td>0.47</td>
<td>0.56</td>
<td>0.51</td>
<td>1400</td>
</tr>
<tr>
<td>CONC</td>
<td>0.34</td>
<td>0.31</td>
<td>0.33</td>
<td>1400</td>
</tr>
<tr>
<td>DEVI</td>
<td>0.62</td>
<td>0.58</td>
<td>0.60</td>
<td>1400</td>
</tr>
<tr>
<td>DISO</td>
<td>0.44</td>
<td>0.50</td>
<td>0.47</td>
<td>1400</td>
</tr>
<tr>
<td>GENE</td>
<td>0.72</td>
<td>0.71</td>
<td>0.72</td>
<td>1400</td>
</tr>
<tr>
<td>GEOG</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>1400</td>
</tr>
<tr>
<td>LIVB</td>
<td>0.58</td>
<td>0.61</td>
<td>0.60</td>
<td>1400</td>
</tr>
<tr>
<td>OBJC</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>1400</td>
</tr>
<tr>
<td>OCCU</td>
<td>0.42</td>
<td>0.36</td>
<td>0.38</td>
<td>1400</td>
</tr>
<tr>
<td>ORGA</td>
<td>0.37</td>
<td>0.32</td>
<td>0.34</td>
<td>1323</td>
</tr>
<tr>
<td>PHEN</td>
<td>0.38</td>
<td>0.39</td>
<td>0.38</td>
<td>1400</td>
</tr>
<tr>
<td>PHYS</td>
<td>0.74</td>
<td>0.82</td>
<td>0.78</td>
<td>1400</td>
</tr>
<tr>
<td>PROC</td>
<td>0.40</td>
<td>0.38</td>
<td>0.39</td>
<td>1400</td>
</tr>
</tbody>
</table>

**Accuracy** 0.48 20923

**Macro Avg** 0.48 20923

**Weighted Avg** 0.48 20923

*Set = balanced data set*

**Epoch = 70, BatchSize = 1024**

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTI</td>
<td>0.34</td>
<td>0.33</td>
<td>0.34</td>
<td>1400</td>
</tr>
<tr>
<td>ANAT</td>
<td>0.50</td>
<td>0.62</td>
<td>0.56</td>
<td>1400</td>
</tr>
<tr>
<td>CHEM</td>
<td>0.49</td>
<td>0.55</td>
<td>0.52</td>
<td>1400</td>
</tr>
<tr>
<td>CONC</td>
<td>0.38</td>
<td>0.28</td>
<td>0.32</td>
<td>1400</td>
</tr>
<tr>
<td>DEVI</td>
<td>0.55</td>
<td>0.59</td>
<td>0.57</td>
<td>1400</td>
</tr>
<tr>
<td>DISO</td>
<td>0.48</td>
<td>0.49</td>
<td>0.48</td>
<td>1400</td>
</tr>
<tr>
<td>GENE</td>
<td>0.69</td>
<td>0.71</td>
<td>0.70</td>
<td>1400</td>
</tr>
<tr>
<td>GEOG</td>
<td>0.36</td>
<td>0.31</td>
<td>0.33</td>
<td>1400</td>
</tr>
<tr>
<td>LIVB</td>
<td>0.59</td>
<td>0.59</td>
<td>0.59</td>
<td>1400</td>
</tr>
<tr>
<td>OBJC</td>
<td>0.35</td>
<td>0.36</td>
<td>0.36</td>
<td>1400</td>
</tr>
<tr>
<td>OCCU</td>
<td>0.40</td>
<td>0.35</td>
<td>0.38</td>
<td>1400</td>
</tr>
<tr>
<td>ORGA</td>
<td>0.37</td>
<td>0.34</td>
<td>0.35</td>
<td>1323</td>
</tr>
<tr>
<td>PHEN</td>
<td>0.38</td>
<td>0.39</td>
<td>0.38</td>
<td>1400</td>
</tr>
<tr>
<td>PHYS</td>
<td>0.75</td>
<td>0.81</td>
<td>0.78</td>
<td>1400</td>
</tr>
<tr>
<td>PROC</td>
<td>0.40</td>
<td>0.39</td>
<td>0.39</td>
<td>1400</td>
</tr>
</tbody>
</table>

**Accuracy** 0.48 20923

**Macro Avg** 0.48 20923

**Weighted Avg** 0.48 20923

*Set = balanced data set*

**LR = 0.0005, Epoch=30, BatchSize=512**

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTI</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>1400</td>
</tr>
<tr>
<td>ANAT</td>
<td>0.60</td>
<td>0.54</td>
<td>0.57</td>
<td>1400</td>
</tr>
<tr>
<td>CHEM</td>
<td>0.52</td>
<td>0.54</td>
<td>0.53</td>
<td>1400</td>
</tr>
<tr>
<td>CONC</td>
<td>0.34</td>
<td>0.33</td>
<td>0.33</td>
<td>1400</td>
</tr>
<tr>
<td>DEVI</td>
<td>0.59</td>
<td>0.57</td>
<td>0.58</td>
<td>1400</td>
</tr>
<tr>
<td>DISO</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>1400</td>
</tr>
<tr>
<td>GENE</td>
<td>0.72</td>
<td>0.73</td>
<td>0.73</td>
<td>1400</td>
</tr>
<tr>
<td>GEOG</td>
<td>0.31</td>
<td>0.37</td>
<td>0.34</td>
<td>1400</td>
</tr>
<tr>
<td>LIVB</td>
<td>0.60</td>
<td>0.61</td>
<td>0.60</td>
<td>1400</td>
</tr>
<tr>
<td>OBJC</td>
<td>0.40</td>
<td>0.34</td>
<td>0.37</td>
<td>1400</td>
</tr>
<tr>
<td>OCCU</td>
<td>0.33</td>
<td>0.38</td>
<td>0.35</td>
<td>1400</td>
</tr>
<tr>
<td>ORGA</td>
<td>0.33</td>
<td>0.36</td>
<td>0.35</td>
<td>1323</td>
</tr>
<tr>
<td>PHEN</td>
<td>0.43</td>
<td>0.35</td>
<td>0.39</td>
<td>1400</td>
</tr>
<tr>
<td>PHYS</td>
<td>0.77</td>
<td>0.81</td>
<td>0.79</td>
<td>1400</td>
</tr>
<tr>
<td>PROC</td>
<td>0.40</td>
<td>0.37</td>
<td>0.39</td>
<td>1400</td>
</tr>
</tbody>
</table>

**Accuracy** 0.48 20923

**Macro Avg** 0.48 20923

**Weighted Avg** 0.48 20923

*Set = balanced data set*
Results pt. 3

Layer 2 Removed
LR = 0.0007, Epoch=30, BatchSize = 512

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTI</td>
<td>0.34</td>
<td>0.37</td>
<td>0.35</td>
<td>1400</td>
</tr>
<tr>
<td>ANAT</td>
<td>0.51</td>
<td>0.64</td>
<td>0.57</td>
<td>1400</td>
</tr>
<tr>
<td>CHEM</td>
<td>0.55</td>
<td>0.48</td>
<td>0.51</td>
<td>1400</td>
</tr>
<tr>
<td>CONC</td>
<td>0.34</td>
<td>0.35</td>
<td>0.35</td>
<td>1400</td>
</tr>
<tr>
<td>DEVI</td>
<td>0.61</td>
<td>0.55</td>
<td>0.58</td>
<td>1400</td>
</tr>
<tr>
<td>DISO</td>
<td>0.50</td>
<td>0.48</td>
<td>0.49</td>
<td>1400</td>
</tr>
<tr>
<td>GENE</td>
<td>0.70</td>
<td>0.71</td>
<td>0.71</td>
<td>1400</td>
</tr>
<tr>
<td>GEOG</td>
<td>0.37</td>
<td>0.32</td>
<td>0.34</td>
<td>1400</td>
</tr>
<tr>
<td>LIVB</td>
<td>0.57</td>
<td>0.62</td>
<td>0.60</td>
<td>1400</td>
</tr>
<tr>
<td>OBIC</td>
<td>0.39</td>
<td>0.35</td>
<td>0.37</td>
<td>1400</td>
</tr>
<tr>
<td>OCCU</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>1400</td>
</tr>
<tr>
<td>ORGA</td>
<td>0.38</td>
<td>0.34</td>
<td>0.36</td>
<td>1323</td>
</tr>
<tr>
<td>PHEN</td>
<td>0.38</td>
<td>0.38</td>
<td>0.38</td>
<td>1400</td>
</tr>
<tr>
<td>PHYS</td>
<td>0.77</td>
<td>0.81</td>
<td>0.79</td>
<td>1400</td>
</tr>
<tr>
<td>PROC</td>
<td>0.36</td>
<td>0.41</td>
<td>0.38</td>
<td>1400</td>
</tr>
</tbody>
</table>

accuracy                           0.48     20923
macro avg       0.48      0.48      0.48     20923
weighted avg   0.48      0.48      0.48     20923

set = balanced data set

Layer 2 removed
LR=0.0005, Epoch=100, BatchSize=256

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.36</td>
<td>0.38</td>
<td>0.37</td>
<td>1400</td>
</tr>
<tr>
<td>1</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
<td>1400</td>
</tr>
<tr>
<td>2</td>
<td>0.48</td>
<td>0.52</td>
<td>0.50</td>
<td>1400</td>
</tr>
<tr>
<td>3</td>
<td>0.35</td>
<td>0.32</td>
<td>0.34</td>
<td>1400</td>
</tr>
<tr>
<td>4</td>
<td>0.60</td>
<td>0.57</td>
<td>0.58</td>
<td>1400</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>0.46</td>
<td>0.48</td>
<td>1400</td>
</tr>
<tr>
<td>6</td>
<td>0.68</td>
<td>0.73</td>
<td>0.70</td>
<td>1400</td>
</tr>
<tr>
<td>7</td>
<td>0.37</td>
<td>0.33</td>
<td>0.35</td>
<td>1400</td>
</tr>
<tr>
<td>8</td>
<td>0.57</td>
<td>0.62</td>
<td>0.59</td>
<td>1400</td>
</tr>
<tr>
<td>9</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>1400</td>
</tr>
<tr>
<td>10</td>
<td>0.34</td>
<td>0.36</td>
<td>0.35</td>
<td>1400</td>
</tr>
<tr>
<td>11</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>1323</td>
</tr>
<tr>
<td>12</td>
<td>0.38</td>
<td>0.38</td>
<td>0.38</td>
<td>1400</td>
</tr>
<tr>
<td>13</td>
<td>0.77</td>
<td>0.83</td>
<td>0.80</td>
<td>1400</td>
</tr>
<tr>
<td>14</td>
<td>0.41</td>
<td>0.35</td>
<td>0.38</td>
<td>1400</td>
</tr>
</tbody>
</table>

accuracy                           0.48     20923
macro avg       0.47      0.48      0.47     20923
weighted avg   0.47      0.48      0.47     20923

set = balanced data set
Future Directions

• Running the model to find the upper limits of the variables I changed for experimentation

• Create a confusion matrix of what atoms got misclassified into which classes

• Binary prediction where the model determines if atoms in the dataset belong to one class, for each class

• Fine tuning the model based on which classes the model isn’t confident about assigning atoms to

• Use BioWordVec with SAP BERT to combine strings with SAB

Epoch = 300, SAB & STR w/ S-B & BWV

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.46</td>
<td>0.62</td>
<td>0.53</td>
<td>400</td>
</tr>
<tr>
<td>1</td>
<td>0.70</td>
<td>0.85</td>
<td>0.77</td>
<td>9363</td>
</tr>
<tr>
<td>2</td>
<td>0.94</td>
<td>0.84</td>
<td>0.89</td>
<td>53736</td>
</tr>
<tr>
<td>3</td>
<td>0.24</td>
<td>0.46</td>
<td>0.32</td>
<td>5013</td>
</tr>
<tr>
<td>4</td>
<td>0.60</td>
<td>0.77</td>
<td>0.67</td>
<td>2499</td>
</tr>
<tr>
<td>5</td>
<td>0.81</td>
<td>0.69</td>
<td>0.74</td>
<td>38546</td>
</tr>
<tr>
<td>6</td>
<td>0.86</td>
<td>0.97</td>
<td>0.91</td>
<td>8932</td>
</tr>
<tr>
<td>7</td>
<td>0.17</td>
<td>0.34</td>
<td>0.23</td>
<td>424</td>
</tr>
<tr>
<td>8</td>
<td>0.96</td>
<td>0.78</td>
<td>0.86</td>
<td>50503</td>
</tr>
<tr>
<td>9</td>
<td>0.18</td>
<td>0.36</td>
<td>0.24</td>
<td>1124</td>
</tr>
<tr>
<td>10</td>
<td>0.16</td>
<td>0.30</td>
<td>0.20</td>
<td>400</td>
</tr>
<tr>
<td>11</td>
<td>0.17</td>
<td>0.28</td>
<td>0.21</td>
<td>400</td>
</tr>
<tr>
<td>12</td>
<td>0.21</td>
<td>0.36</td>
<td>0.27</td>
<td>800</td>
</tr>
<tr>
<td>13</td>
<td>0.93</td>
<td>0.97</td>
<td>0.95</td>
<td>12588</td>
</tr>
<tr>
<td>14</td>
<td>0.53</td>
<td>0.81</td>
<td>0.64</td>
<td>18699</td>
</tr>
</tbody>
</table>

- accuracy                           0.79    203427
- macro avg 0.53 0.63 0.56 203427
- weighted avg 0.84 0.79 0.80 203427
set = 10 % subset
References


- UMLS® Reference Manual [Internet]. Bethesda (MD): National Library of Medicine (US); 2009 Sep-. Table 1. [Concept Names and Sources (File = MRCONSO.RRF)]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9685/table/ch03.T_concept_names_and_sources_file_mrr/


Thank You