**Abstract**

Mapping and translating professional but arcane clinical jargons to consumer language is essential to improve the patient-clinician communication. We utilized the embeddings alignment method for the word mapping between unparalleled clinical professional and consumer language embeddings.

To map semantically similar words in two different word embeddings, we independently trained word embeddings on the professional and layman clinical corpora. Then, we aligned the embeddings by the Procrustes and adversarial algorithms with refinement. We evaluated the quality of the alignment through the similar words retrieval both by the model precision and human judgment.

The Procrustes algorithm can be performed for the alignment, whereas adversarial training with refinement may find some relations between professional and consumer language embeddings.

**Background**

**Motivation**

- Patient-clinician communication
- Clinical document
- Huge information gap between professionals and consumers
- On floor pt found to be hypoxic on D2 4LNC O2 sats 85 %, CXR c/w pulm edema, she was given 40mg IV x 2, nebs, and put out 1.5 L UOP, she was also put on a NRB with improvement in O2 Sats to 95 %

**Affecting Clinical Decision Making**

- Breast cancer/lesion/abnormal cells [Omer 2013]
- PCOS/hormone imbalance [Copp 2017]
- Result in...
- Over-treatment / over-diagnosis / defensive medicine

**Current Solutions**

- UMLS/Consumer health vocabulary (CHV)
- Synonym replacement
- Explanation insertion
- Problem - expert efforts to manually build the dictionary, which is hard to be generalized and scalable
- Pattern-based mining with Wikipedia corpus [Vydawaharan 2014]
- Problem - not specific for professional and consumer languages

**Proposed Approach**

- Unsupervised word vector representation [Mikolov 2013, Bojanowski 2017]
- Embeddings alignment [Conneau 2018, Chung 2018, Lample 2018]

---

**Learning Word Embeddings**

- **Word-level** (word2vec) [Mikolov 2013]
- **Subword-level** (fasttext) [Bojanowski 2017]
- Different window sizes (k=3, 5)

**Alignment - Procrustes Algorithm**

- Constructing the synthetic mapping dictionary to learn a linear mapping matrix between the two embedding spaces

\[
W^* = \text{argmin}_W |WX - YY^T|_F^2
\]

\[
W^* = \text{argmin}_W \|UW - VY^T\|_F^2, \text{ where } U^T V = \text{SVD}(YX^T)
\]

**Alignment - Adversarial Training**

- Making the aligned embeddings indistinguishable

\[
L_{\phi}(W) = -\frac{1}{k} \sum_{y=1}^{k} \log P_{\phi}(y|W) - \frac{1}{k} \sum_{x=1}^{k} \log P_{\phi}(W|x)
\]

\[
L_{W}(\phi(y|x)) = -\frac{1}{k} \sum_{y=1}^{k} \log P_{\phi}(y|W) + \frac{1}{k} \sum_{y=1}^{k} \log P_{\phi}(y|W)
\]

**Methods**

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Embedding</th>
<th>Window</th>
<th>P@5</th>
<th>P@5</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIMIC-P</td>
<td>MIMIC-C</td>
<td>word</td>
<td>5/5</td>
<td>0.19</td>
<td>0.42</td>
<td>0.54</td>
</tr>
<tr>
<td>MIMIC-P</td>
<td>MIMIC-C</td>
<td>subword</td>
<td>5/5</td>
<td>0.27</td>
<td>0.57</td>
<td>0.78</td>
</tr>
<tr>
<td>PMC-Publicmed</td>
<td>MIMIC-C</td>
<td>word</td>
<td>30/3</td>
<td>0.26</td>
<td>0.40</td>
<td>0.44</td>
</tr>
<tr>
<td>PMC-Publicmed</td>
<td>MIMIC-C</td>
<td>subword</td>
<td>30/3</td>
<td>0.18</td>
<td>0.39</td>
<td>0.44</td>
</tr>
<tr>
<td>PMC-Publicmed</td>
<td>Wikipedia</td>
<td>word</td>
<td>30/10</td>
<td>0.23</td>
<td>0.43</td>
<td>0.49</td>
</tr>
<tr>
<td>PMC-Publicmed</td>
<td>Wikipedia</td>
<td>subword</td>
<td>30/10</td>
<td>0.23</td>
<td>0.43</td>
<td>0.49</td>
</tr>
</tbody>
</table>

**Data Source**

- 59,654 free-text discharge summaries in MIMIC-III version 1.4.
- Professional corpus (443,585 sentences / 26,333 vocabularies)
- History of present illness
- Brief hospital course
- Consumer corpus (73,349 sentences / 6,752 vocabularies)
- Discharge instruction
- Followup instruction
- Source and target corpora are not parallel
- Overtreatment / overdiagnosis / defensive medicine

**Conclusion**

We demonstrate the capability of embeddings alignment for mapping unparalleled clinical professional and consumer languages in word-level, without the knowledge and supervision from biomedical ontologies and dictionaries, and just use the minimal supervision using the identical strings across corpora. We found that the Procrustes algorithm with anchors approach with the subword-level word embeddings trained on clinical narrative texts, rather than larger general corpus, outperformed the other combinations.

The aligned embeddings learned from the adversarial training approach reveal the relation between professional and consumer anatomy-related terms. Further investigations include exploring larger datasets and extending word-level to concept-level embeddings.