A Sequence Labeling Approach to Deriving Word Variants

Jennifer D’Souza
Human Language Technology Research Institute
University of Texas at Dallas, Richardson, TX 75083-0688
jennifer.l.dsouza@utdallas.edu

Abstract
This paper describes a learning-based approach for automatic derivation of word variant forms by the suffixation process. We employ the sequence labeling technique, which entails learning when to preserve, delete, substitute, or add a letter to form a new word from a given word. The features used by the learner are based on characters, phonetics, and hyphenation positions of the given word. To ensure that our system is robust to word variants that can arise from different forms of a root word, we generate multiple variant hypotheses for each word based on the sequence labeler’s prediction. We then filter out ill-formed predictions, and create clusters of word variants by merging together a word and its predicted variants with other words and their predicted variants provided the groups share a word in common. Our results show that this learning-based approach is feasible for the task and warrants further exploration.

Introduction
An automatic word variant derivation component as part of lexicon builder tools assuages the otherwise tedious manual effort of populating variants of each word in the lexicon. Related work (Lu et al. 2012) in this direction, have developed an automatic word variant derivation component by following a purely rule-based approach. In general, while rule-based approaches are credited as being highly precise, since their applicability is restricted to only seen patterns, their coverage remains a limiting factor. As language is ever-evolving, especially in certain domains like the medical domain, incorporating machine learning-based methods either as a standalone or in a hybrid architecture together with rule-based approaches seems a worthwhile next step to enable further development of this task.

The word variant derivation task by suffixation can be described as follows. Given a word, to determine its unchangeable beginning portion (loosely, the base), identifying any remaining ending as replaceable (loosely, the suffix), and potential candidate suffixes for replacement or for appending to the base resulting in valid derived words (e.g., happ-iness from happ-y). In this work, we propose a learning-based approach to the task of word variant derivation by suffixation. We present our work in the following sections beginning with the experimental data creation process, followed by details on our adopted approach, and concluding with an evaluation of the approach.

Data
We created a bag-of-words dataset for this project using a corpus of about 400,000 clinical notes. The bag-of-words was then organized into word groups so that each group contained all the words that were derivational variants of each other. To do this, WordNet (Miller 1995) was queried for the synsets of each word in the bag. On obtaining the synsets of a word, each synset-form was queried for its derivational word variant list, which if not empty was retrieved. Further pruning of the bag-of-words was carried out to remove all words without any derivational variants.

Our final dataset, organized as groups of words, contained 12,057 unique words and a total of 4609 word groups. Three example groups from WordNet are listed below.

(1) pharmacy, pharmaceutic, pharmacist, pharmaceutical
(2) pathological, pathologist, pathology, pathologic
(3) digestion, digest, digester, digestive, digestible

Our Learning-Based Approach
Our preliminary approach to automatic word variant derivation by suffixation is similar to the nonstandard word normalization tasks described in Pennell and Liu (2010) and Liu et al. (2011).

Word Variant Derivation by Sequence Learning
We treat the problem as a multi-class classification task. For each letter, a decision is made whether or not to preserve, delete, or substitute the letter; and if the variant word is hypothesized to have more letters, to add a new letter to the original word. This transformation process of a word to its variants is learned automatically through a sequence labeling framework that integrates character-, phonetic-, and hyphenation information. We used DirecTL+ (Jiampoja-
We conducted three different evaluations of our approach: 1) Exact - shows performance in predicting derivational word variant groups in the original dataset exactly; 2) Entire - reflects performance of predicting the original dataset groups entirely, which means the predicted groups can contain more derived words; and 3) Partial - shows performance in predicting original dataset groups at least partially. Performances are measured as recall (R), precision (P), and F-score (F). The “All groups” column in Table 1 shows these evaluation results. In addition, we evaluated the sequence learner’s performance on a subset of our original data containing only those groups with words that were entirely new to the model. A word was categorized as entirely new if its three-letter prefix did not match any of the training data words three-letter prefixes. The “New groups” column in Table 1 shows these results.

Thus our proposed learning-based approach for word variant derivation shows itself a suitable next step to the task development in that: it learns the base word for suffixation; it learns the replaceable suffixes in a word; it learns the mapping between replaceable suffixes and eligible suffixes for replacement; and it is not restricted by a single direction of derivation.

**Results and Conclusion**

We conducted three different evaluations of our approach:

<table>
<thead>
<tr>
<th>All groups</th>
<th>New groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>77.7</td>
</tr>
<tr>
<td>P</td>
<td>78.5</td>
</tr>
<tr>
<td>F</td>
<td>86.6</td>
</tr>
</tbody>
</table>

**References**


SNOEMED CT or in the dataset.

---


Predictions are filtered out as ill-formed if not found in

Figure 1: Training data examples of (word, variant) pairs.