Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

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 hypothesis 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[1] 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.[2] We used DirecTL+ (Jiampoja-

[1] Filtered to retain only the words found in the SNOMED clinical terminology, WordNet, or in the online Webster's dictionary.

[2] Additional details are available at http://www.hlt.utdallas.edu/~jld082000/derivation/

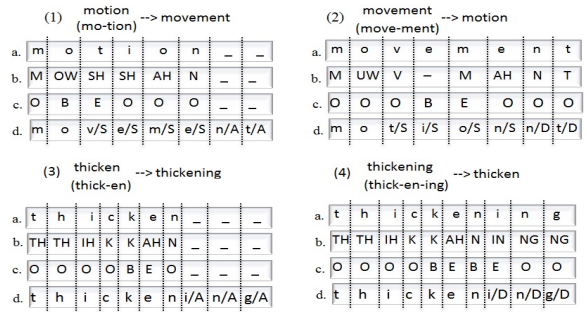

Figure 1: Training data examples of (word, variant) pairs.

marn, Cherry, and Kondrak 2010) for phonetic features, and Liang's (1983) algorithm for hyphenation features.

Figure 1 shows some example (word, variant word) pairs in our training data. It depicts the feature representations for each word: 1. its alphabet sequence (see row a); 2. spelled-out alphabet-aligned phoneme pronunciation (see row b); and 3. the given word's valid hyphenation points as begin [B], end [E], and outside [O] (see row c). In the figure, row d, shows the labels (classes) applied to each letter in the given word for deriving the variant word: one of the 26 alphabets; or an alphabet marked with an S for substitution, D for deletion, or A for addition.

The first feature was intended to capture suffixes for replacement (e.g., "-ion"), while the second feature was intended to qualify the first feature (e.g., *n* with a spelled pronunciation as "NG" almost certainly represents a character which can be deleted or substituted to form a new word, but *n* with a spelled pronunciation "N" varies in its role). And the third hyphenation feature was used since in many cases the alphabets following hyphenation breakpoints serve as valid suffixes (e.g., "-ening", or "-ing" from *thick-en-ing*).

We employed CRF++[3] as the sequence learner. To train the model, we used a subset of the dataset described in the earlier section containing 3885 unique words and a total of 1500 word groups. Training instances were created as follows. From each group of words that were derivational variants of each other, we created one instance by pairing a word with another word in the same group. In this way, training instances comprised all possible pairings of words within the group. Note that we fix no criteria for the direction of derivation of a word from another. Thus the sequence learner has more possibilities to explore in finding the most frequent patterns for suffixation.

The trained model is then applied on test instances created from words in the remaining dataset. During testing, to facilitate the model's ability to generate variants longer than the given word, several test instances were created for a word starting with the word itself, and a new instance each time the word is appended with a blank (like (1) and (3) in Figure 1) for each unique word length longer than itself and up to the longest in the corpus. For each test instance, we consider the predicted sequence with the highest marginal probability as its variant. Thus as a result of automatic suffixation by the CRF model, each word may have zero[4] or more derived word predictions after the testing phase. We

[3]Available from http://crfpp.sourceforge.net

[4]Predictions are filtered out as ill-formed if not found in SNOMED CT or in the dataset.

	All groups			New groups		
	Exact	Entire	Partial	Exact	Entire	Partial
R	53.1	77.0	90.4	54.8	68.5	82.8
P	60.9	78.5	86.9	40.7	48.8	56.3
F	56.7	77.7	88.6	46.7	57.0	67.0

Table 1: Results for grouping derivational forms of words by automatic suffixation

then merged together each test instance word and its variants with other test instance words and their variants provided the groups shared a word in common. In this way, our results were organized in the same manner as the original dataset in the form of groups containing derivational variants of words.

Results and Conclusion

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 atleast 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.

References

Jiampojamarn, S.; Cherry, C.; and Kondrak, G. 2010. Integrating joint n-gram features into a discriminative training framework. In *Proceedings of NAACL-2010*.

Liang, F. M. 1983. *Word hyphenation by computer*. Department of Computer Science, Stanford University.

Liu, F.; Weng, F.; Wang, B.; and Liu, Y. 2011. Insertion, deletion, or substitution?: normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th ACL*, 71–76.

Lu, C. J.; McCreedy, L.; Tormey, D.; and Browne, A. C. 2012. A systematic approach for medical language processing: Generating derivational variants. *IT Professional* 14(3):36–42.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Pennell, D. L., and Liu, Y. 2010. Normalization of text messages for text-to-speech. In *ICASSP*, 4842–4845.