Improving Automated Prediction of English Lexical Blends Through the Use of Observable Linguistic Features

Jarem Saunders
M.A. Student, Department of Linguistics
University of North Carolina at Chapel Hill
jsaunders1@unc.edu   jarem.saunders@gmail.com

Abstract

The process of lexical blending is difficult to reliably predict. This difficulty has been shown by machine learning approaches in blend modeling, including attempts using then state-of-the-art LSTM deep neural networks trained on character embeddings, which were able to predict lexical blends given the ordered constituent words in less than half of cases, at maximum. This project introduces a novel model architecture which dramatically increases the correct prediction rates for lexical blends, using only Polynomial regression and Random Forest models. This is achieved by generating multiple possible blend candidates for each input word pairing and evaluating them based on observable linguistic features. The success of this model architecture illustrates the potential usefulness of observable linguistic features for problems that elude more advanced models which utilize only features discovered in the latent space.

1 Introduction

1.1 Descriptive Research on Lexical Blends

Lexical blends have long been noted as a linguistic phenomenon with little consistent predictability. Researchers have described many different factors which affect how much of two given input words will be preserved in the resulting blend. This is often described in terms of the "switchpoint," or point at which each input word is truncated.

Factors described in the literature include a tendency for words to split at syllable constituent boundaries (Gries 2012, Kelly 2009), an observation that blends tend to match the length of the 2nd input word (Bat-El 2006), and a finding that the prosodic structure (Arndt-Lappe & Plag 2013). None of these noted tendencies or a combination thereof has thus far been used to create a predictive model of blending.

1.2 Predictive Models of Lexical Blends

Researchers who have used data-driven methods to model blending have instead opted for the use phoneme-by-phoneme insertion and deletion counts or the use of character embeddings. The former of these approaches was used by Deri & Knight (2015) as part of a multi-tape FST model that used grapheme-phoneme alignments to train the model on transformations to the phoneme sequence and produce the correct orthographic output, achieving a maximum of 45.75% correct blend predictions.

Gangal et. al. (2017) used the latter approach, training a then state-of-the-art LSTM deep neural network on character embeddings to attempt to generate English-like blends. This was shown to improve the rate of correct predictions to 48.75%, and found that the best performing models entertained multiple blend candidates and selected the most probable form, described as "exhaustive generation", rather than using greedy decoding from the vector space. Both of these models often produced sequences which were phototactically invalid, though sometimes these were orthographically plausible.

Because of the overall limited success of the models, including only a small increase in performance between the models despite a large increase in model complexity, we have developed an alternative model architecture which utilizes the same grapheme/phoneme alignment system as Deri & Knight and the exhaustive generation strategy laid out by Gangal et. al., but uses

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 93–97
July 14, 2023 ©2023 Association for Computational Linguistics

linguistically-motivated features which were directly observable from the input forms. The use of linguistically-informed feature spaces was shown to improve performance in blend prediction using a modified form of the Gangal et. al. LSTM architecture, though improvements were once again quite modest (Kulkarni & Wang 2018). This paper proposes a more dramatic change in architecture which uses a novel feature set based primarily on the descriptive blend characteristics of Arndt-Lappe & Plag (2013).

For the purpose of this analysis, we constraint the blend structures entertained to only be those that follow typical English blend formation patterns by keeping some initial portion of the first word and some final portion of the second word. This model architecture was applied to 3 different corpora of lexical blends and was compared to previous model performance on each corpus, when applicable.

2 Methods

The model architecture laid out in this paper included the following elements:

- A component to generate all plausible blend candidates from the two input words and extract linguistically-based feature values using grapheme/phoneme alignments and syllable structure information.
- A component which uses the extracted features to calculate the probability of being a valid blend for each blend candidate.
- A feature to select the most probable candidate from each input word pairing.

The generation process was performed by iteratively creating prefixes from the first input word and suffixes from the second input word using grapheme-phoneme alignments, such that the substring consisted of a contiguous sequence of phonemes and their corresponding graphemes. Each prefix was then concatenated with each suffix to produce the full candidate set for each input word pair.

Feature values were calculated for each candidate using a set of phonemically-defined features, modified from Arndt-Lappe & Plag (2013). Labels were assigned to each candidate based on whether the graphemes of the candidate matched the desired blend output. Candidates with feature values which were identical to a candidate already in the feature set were removed.

Feature names	Description
W1/W2 length	number of syllables
Candidate length	number syllables
Medial overlap	whether candidate has contiguous phonemes shared by prefix/suffix
W1/W2 left/right edge to primary stress	number of syllables from input word edges to primary stress
W1 left edge to switchpoint	number of syllables from W1 left edge to switchpoint
W2 left edge to switchpoint	number of syllables from W2 right edge to switchpoint
Switchpoint syllable bound	whether switchpoint occurs at onset, nucleus, or coda boundary, or not at boundary
W1/W2 primary stress preserved	whether candidate preserves primary stress of input(s)
W1/W2 segments preserved	proportion of segments from each input preserved in candidate
W1/W2 syllables preserved	proportion of syllable nuclei from each input preserved in candidate
Switchpoint at W2 primary stress syllable	whether switchpoint in W2 falls within primary syllable bearing primary stress

Table 1: Complete set of linguistically-based model features utilized in trials

Given the feature values for all candidates, we used Random Forest classifiers and Polynomial regression models to learn probabilities for each candidate based on the extracted feature values. Rather than assign each data instance to a class, probabilities are retained for each candidate so that the candidate with the maximum probability can be selected. Random Forest and Polynomial regression were chosen for this experiment because they are easier to train and interpret than deep neural approaches.

Finally, a selection component was used to find the candidate from each input word pairing with the greatest probability of being a valid blend of English. This candidate was then chosen as the model's predicted output for that input word pairing.

2.1 Feature Set

The model used features which were modeled after the most relevant cues for blends discussed in previous linguistic literature on blend formation and structure. Among these are features that track whether the switchpoint aligns with syllable structure boundaries, the proportion of the input word that is preserved, and which word (if any) has it's stress patterns preserved.

In addition to these phonological features derived directly from the phoneme representations of input words, the model uses the phonotactic markedness score calculated by the BLICK phonotactic learner, which returns a score to indicate how well a sequence of phonemes follows English phonotactics (Hayes 2012). This is expressed as a sum of weighted violations of MaxEnt grammar constraints learned from a large sample of words from the CMU pronouncing dictionary (Hayes 2008). This feature was included to improve the phonotactic plausibility of output candidates. Specific names and descriptions for all model features are given in Table 1.

3 Experiments

The specific trials the model was used for are given here, along with the datasets utilized in training/testing for those trials.

3.1 Datasets

Three separate corpora were used in training and testing the model. The first two corpora were those used in the previous machine learning models of blending, Deri & Knight (2015) and Gangal et. al. (2017), respectively. These were both acquired through online resources such as Wikipedia, Wiktionary, and Urban Dictionary. The final corpus used comes from Shaw (2014) and is a curation of an earlier dictionary assembled by Thurner (1993). After filtering to meet the project design, 322 blends were used in trials of the Deri & Knight corpus, 1092 were used from Gangal et. al., and 1096 were used from Shaw.

3.2 Trials Performed

For each corpus, the model architecture was tested and evaluated using three different learners: LASSO regression, 2nd order Polynomial regression (with interaction terms), and Random Forest classifiers. Due to high collinearity among the feature set, we selected subset of the model features was selected to minimize correlations and maximize coefficient values by removing measures of syllable proportion, word length, and syllable distances to the switchpoint from the feature set. Each learner was trained once using the full feature set and once using the manually selected subset of features. Each trial was validated using 10-fold cross validation.

Learner	**Features**	**Correct**	**Edit dist.**
LASSO	Full	56.13%	1.05
Polynomial	Full	64.42%	0.72
RF	Full	60.39%	0.81
LASSO	Subset	55.21%	1.09
Polynomial	Subset	63.83%	0.79
RF	Subset	60.39%	0.83
Previous benchmark		45.39%	1.59

Table 2: Model Performance on D.&K. Corpus

Learner	**Features**	**Correct**	**Edit dist.**
LASSO	Full	47.72%	1.36
Polynomial	Full	59.51%	0.89
RF	Full	57.32%	0.89
LASSO	Subset	46.17%	1.43
Polynomial	Subset	54.21%	1.06
RF	Subset	57.32%	0.95
Previous benchmark		48.75%	1.12

Table 3: Model Performance on G. et. al. Corpus

Learner	**Features**	**Correct**	**Edit dist.**
LASSO	Full	66.09	0.93
Polynomial	Full	74.13%	0.58
RF	Full	73.96%	0.58
LASSO	Subset	64.17%	0.98
Polynomial	Subset	71.67%	0.66
RF	Subset	72.58%	0.65

Table 4: Model Performance on Shaw Corpus

4 Results

Model predictions were evaluated on the average percentage of blends correctly predicted and the average Levenshtein edit distance between the predicted output form and the correct blend form.

4.1 Quantitative Results

Models trained and tested on the Deri & Knight corpus outperformed the benchmark set by the multi-tape FST on both measures of model performance for every variation of the model. For the Gangal et. al. corpus, only the variation of the model using the LASSO regression learner failed to outperform the benchmark set by the LSTM model using character embeddings. The Shaw corpus demonstrated the highest performance of any model, with an average of 74.13% of blends correctly predicted with the best performing model trained on this corpus.

For all corpora, the model variations that used Polynomial regression learners outperformed all others, and models using full data set outperformed those with the manually curated subset, in spite of the high collinearity of the dataset.

A comparison of the highest performing models to date for all corpora is given in Figure 1.

4.2 Qualitative Results

Qualitative error analysis shows that the best performing model across all corpora, the Polynomial regression with full features trained on the Shaw corpus, tends to over-preserve phonemic material from both input words, rather than over-delete. In a random sample of 100 instances in which this model selected an incorrect candidate, 40 of them preserved too many contiguous segments from the first word, compared to 15 instances in which the output candidate had a sequence from the first word which was too short. Similarly, the sample demonstrated that 33 candidates had over-preserved segmental material from the second input word, compared to 18 instances in which too many segments of the word were deleted. In general, this resulted in a greater number of candidates that were longer than the desired output than candidates that were too short.

5 Discussion

5.1 Usefulness of Observable Features

Results from the trials we have conducted so far provide a compelling argument for the potential usefulness of observable linguistic features in the generation of lexical blends. This, in turn, may provide a framework for dealing with similar linguistic processes which exhibit some degree of unpredictability or are infrequently attested in natural language text data and accordingly are difficult for models which rely on features obtained in the latent space.

Little work has been done to date to tune hyperparameters or optimize the feature set used by the models. Future research into these areas could lead further improvements in the prediction rates that has already gained by using this model architecture, including research into measures to reduce the apparent model bias toward longer candidates.

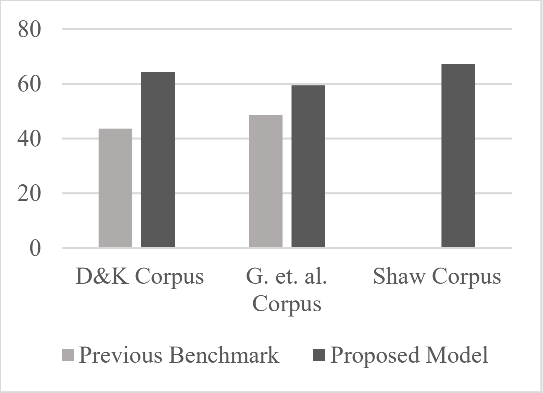

Figure 1: Maximum Model Correct Prediction Rates by Corpus

5.2 Potential Linguistic Applications

The architecture may also be useful for testing hypotheses about blend generation. Because the learners used in this model architecture are more interpretable than neural networks, the actual feature weights and decision tree splits used by the model can be directly examined and can be used as a datapoint in evaluating the relative importance of different factors that affect blend formation. Given the fact that there are often many possible blends that speakers can produce from an input word pair before it enters the lexicon (Gries 2012), testing the model's performance on novel blend forms and comparing it to blends produced by human speakers would be the most informative way to test how well this model truly does at replicating human-like blend generation behavior.

Such a trial would also be informative in comparing this methodology against modern large language models, as it provides a chance to use genuinely held-out data to evaluate them. One drawback of this architecture is its lack of generalizability to low resource languages.

5.3 Limitations of the Model Architecture

While this methodology does not require the large amount of text data utilized by more advanced models, it does depend on access to grapheme/phoneme alignment information for all input words. This does limit the usefulness of the model for languages with little linguistically-tagged data available, though the success of the small Deri & Knight corpus does indicate that the model architecture can be made to function effectively with a limited amount of annotated data.

References

Arndt-Lappe, S., & Plag, I. (2013). The role of prosodic structure in the formation of English blends. English Language & Linguistics, 17(3), 537-563.

Deri, A., & Knight, K. (2015). How to make a frenemy: Multitape FSTs for portmanteau generation. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 206-210).

Gangal, V., Jhamtani, H., Neubig, G., Hovy, E., & Nyberg, E. 2017. Charmanteau: Character embedding models for portmanteau creation. arXiv preprint arXiv:1707.01176.

Gries, S. T. 2004. Shouldnt it be breakfunch? A quantitative analysis of blend structure in English.

Gries, S. T. 2006. Cognitive determinants of subtractive word formation: A corpus-based perspective.

Gries, S. T. 2012. Quantitative corpus data on blend formation: Psycho-and cognitive-linguistic perspectives. Cross-disciplinary perspectives on lexical blending, 252, 145.

Hayes, B. 2012. BLICK: a phonotactic probability calculator (manual).

Hayes, B., & Wilson, C. 2008. A maximum entropy model of phonotactics and phonotactic learning. Linguistic inquiry, 39(3), 379-440.

Kelly, M. H. 1998. To “brunch” or to “brench”: Some aspects of blend structure.

Kubozono, H. 1990. Phonological constraints on blending in English as a case for phonology-morphology interface. Yearbook of morphology, 3, 1-20.

Kulkarni, V., & Wang, W. Y. (2018, June). Simple models for word formation in slang. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 1424-1434).

Shaw, K. E., White, A. M., Moreton, E., & Monrose, F. 2014. Emergent faithfulness to morphological and semantic heads in lexical blends. In Proceedings of the annual meetings on phonology (Vol. 1, No. 1).

Thurner, Dick. 1993. Portmanteau dictionary: Blend words in the English language, including trademarks and brand names. Jefferson, NC: McFarland & Co.