

Pairing Orthographically Variant Literary Words to Standard Equivalents Using Neural Edit Distance Models

Craig Messner

Center for Digital Humanities
Johns Hopkins University
cmessne4@jhu.edu

Tom Lippincott

Center for Digital Humanities
Johns Hopkins University
tom.lippincott@jhu.edu

Abstract

We present a novel corpus consisting of orthographically variant words found in works of 19th century U.S. literature annotated with their corresponding "standard" word pair. We train a set of neural edit distance models to pair these variants with their standard forms, and compare the performance of these models to the performance of a set of neural edit distance models trained on a corpus of orthographic errors made by L2 English learners. Finally, we analyze the relative performance of these models in the light of different negative training sample generation strategies, and offer concluding remarks on the unique challenge literary orthographic variation poses to string pairing methodologies.

1 Introduction

Using orthographic information to pair similar strings from a list of variants has a number of uses, from spelling correction to cognate detection. Beyond character level similarity, what it means to be a "good" neighbor to a given source string might entail phonetic similarity (in the case of many misspellings), some sort of cognitive proximity (in certain "point" misspellings, say "k" for "c") or it may reflect a shared linguistic history (in the case of cognates). Using orthographic information to achieve a string-pair ranking that incorporates these axes of meaning requires that the orthography of a string also captures this other meaningful dimension, to a greater or lesser degree. We introduce the domain of "literary orthographic variants," a set of orthographic modifications motivated by literary aesthetic concerns instead of purely linguistic or cognitive principle, and posit that the orthographic results of these unique motivations necessitates re-evaluating modeling approaches that have proven successful in more linguistically motivated domains (such as the aforementioned cognate detection and spelling correction tasks). We evaluate this claim using string pairing, a task where a

model compares two strings and outputs a probability that they are a match. These strings can be considered a match if their pair probability exceeds a certain threshold. Furthermore, these probabilities can also be used to rank a set of possible string pairs. We obtain these probabilities using a neural edit distance model architecture, an approach that has proven effective at pairing cognate words. Specifically we train neural edit distances models on a corpus of nonliterary orthographic variants produced by L2 English learners and a novel corpus of literary variants in order to offer the following contributions:

- The aforementioned novel corpus of literary orthographic variants that can support training and evaluation of string-pairing models
- Analysis of this corpus and how its specific set of challenges differ from datasets of orthographic variants that are not derived from literary sources
- Evaluation of the impact of negative example generation strategies on model performance across different domains
- Initial steps towards a general system able to pair literary orthographic variants to their standard forms

2 Background

2.1 Literary Orthographic Variation

While orthographic variation is often framed by deviance from an accepted standard, it has also been used in a literary context as a vehicle of meaning. This technique is notably prevalent in the literature of the 19th century United States, where it often served to identify a particular character as belonging to a certain race, class, gender or region (Ives, 1971) (Jones, 1999). Buoyed by English orthography's highly redundant nature (Shannon, 1951) the

presence of topic-specific surrounding context, and the desire to have the variation itself be meaningful (perhaps by using a particular system of orthography that signifies a certain subject position) literary orthographic variants are typically more extreme and more obscurely motivated than those produced as the result of misspellings or other similar processes.

2.2 String Pairing Using Edit Distance Methods

Edit distance measures, most commonly the Levenshtein Distance (Levenshtein et al., 1966) have been used to rank variant-standard token pairs. More recently, statistical edit distance (Ristad and Yianilos, 1998) and neural edit distance (Libovický and Fraser, 2022) have allowed edit distance to be learned empirically from data. While statistical edit distance learns a single distribution of edit operations over paired strings, neural edit distance uses a differentiable version of the expectation maximization (EM) algorithm as a loss function for a neural model. This allows neural edit distance to learn edit operation probabilities from contextual embeddings. (Libovický and Fraser, 2022) train a neural edit distance string pairing model that employs RNN learned embeddings and randomly generated negative samples in order to achieve state of the art performance on a cognate detection task (Rama et al., 2018).

3 Methods and Materials

3.1 Project Gutenberg Corpus¹

We first use the Project Gutenberg (GB) catalog file² to subset the full collection to English texts produced by authors living in the 19th century. We then limit this set to those works identified as part of the Library of Congress "PS" (American Literature) classification group. We tokenize each of this subset of texts and split into sentences before automatically identifying possible orthovariant tokens using a variety of criteria, including:

- Presence of numeric characters
- Presence of capitalization
- Presence of candidate token in WordNet (Miller, 1995) or the Brown Corpus (Francis and Kucera, 1964)

¹The full corpus is available [here](#)

²available [here](#)

We sampled sentences with possible orthovariant tokens randomly, and Author 1 provided standard token annotations for the tokens deemed actually variant. The final corpus consists of 3058 variant tokens paired with their standard variants and their sentence-level context.

3.2 FCE Corpus

The Cambridge Learner First Certificate in English (FCE) corpus is comprised of short narratives produced by English as a second language (ESL) learners (Yannakoudakis et al., 2011). The corpus includes hand tagged corrections for a variety of observed linguistic errors. We subsetting the corpus to only include errors with a possible orthographic component, indicated by the "S" class of error codes (Nicholls, 2003). This resulted in a subset of 4757 samples.

3.3 Empirical Characterization of Corpora

Corpus	1LD%	2LD%	3LD%	4+LD%
FCE	74.1	20.9	3.2	1.8
Gutenberg	43.8	28.9	17.2	10.1

Table 1: Levenshtein distances of standard and nonstandard tokens in tagged samples, expressed as percentage.

Consistent with our hypothesis about the differences between literary and nonliterary orthovariants, Table 1 demonstrates that the nonstandard tokens found in GB tend to be more distant from their "standard" pairings. This empirically demonstrates at least one axis of difference between the GB corpus and corpora commonly used to evaluate approaches to string pairing, alignment and ranking.

3.4 Experiment 1: Neural Edit Distance String Pairing for Candidate Filtering³

We train a neural edit distance model on a string pairing task and empirically derive a probability threshold in order to separate likely variant/standard token pairs from unlikely pairs. We generate negative samples by pairing variant observed tokens with tokens drawn from Brown using the following methods:

1. Random: n randomly selected known false tokens sourced from Brown

³Code for the following experiments is available at <https://github.com/comp-int-hum/edit-distance-ml>

		F-FCE	F-GB	MRR-FCE	MRR-GB
LD	10	0.81	0.69	0.40	0.29
	20	0.79	0.66	0.64	0.34
	30	0.76	0.60	0.67	0.41
	50	0.72	0.56	0.63	0.44
mixed	10	0.84	0.72	0.59	0.52
	20	0.81	0.67	0.65	0.57
	30	0.79	0.68	0.68	0.56
	50	0.77	0.62	0.67	0.62
random	10	0.97	0.93	0.61	0.47
	20	0.97	0.93	0.65	0.53
	30	0.96	0.90	0.69	0.52
	50	0.94	0.87	0.70	0.50

Table 2: The blue highlighted cell is the best performing model trained on the FCE corpus, red is the best performing trained on the GB corpus. F scores indicate each model’s ability to distinguish true and false string pairs, MRR scores indicates the ability of each model to rank a set of standard-variant pairs generated using Brown

- LD: n lowest LD from source variant known false tokens
- Mixed: $n/2$ Random process tokens and $n/2$ LD process tokens

We perform this procedure for n of 10, 20, 30 and 50. We split the data into test, train and validation sets, each containing a (necessarily unique) admixture of known positive and known negative generated pairs. Following the method of (Libovický and Fraser, 2022) we generate a match probability threshold for each model during training by adjusting it to maximize evaluation F1 score, and then evaluated each model’s ability to distinguish true and false token pairings also using F1 score.

3.5 Experiment 2: Neural Edit Distance String Pairing for Pair Prediction

Leaving aside negative generate pairs, we pair each known true source token in a given test set with all of the tokens found in Brown. We then employ the models trained in Experiment 1 to rank the probability of each source-Brown pairing being a true pair. We evaluate the accuracy of these rankings using mean reciprocal rank (MRR).

4 Results and Analysis

Results of the experiments can be found in Table 2. The F-score of each experiment necessarily depends on the unique set of negatives generated by a given count (10, 20 etc.) and technique (LD,

random, mixed). As it might be expected, models given the more difficult task, in whole or part, of distinguishing low LD variants perform worse. However, the inclusion of these difficult pairs seem to benefit GB’s performance when it comes to overall pair ranking. These MRR scores (columns MRR-FCE and MRR-GB) are produced using only Brown and a stable test set of known positive pairs in each given corpus, and thus form the basis of our comparison.

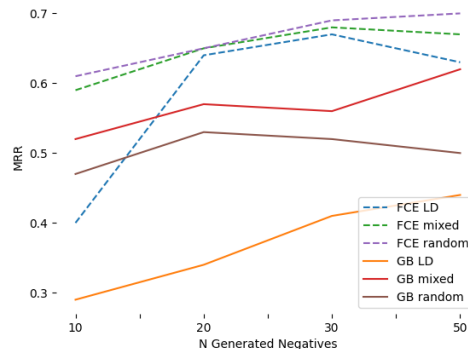


Figure 1: MRR by N Generated Negatives

Figure 1 shows that for all n of negative samples, the models trained on FCE perform best when provided negative samples generated by the random process. The models employing close LD negatives performed uniformly the worst. This is somewhat the opposite of our expected result. The negative signal of incorrect close LD examples would on

its face seem particularly useful given the near LD nature of FCE’s spelling and usage errors, as distinguishing between the candidates in the near LD neighborhood of a variant token becomes imperative.

On the other hand, models trained on GB perform uniformly the best when provided negatives generated by the mixed strategy, a combination of random and close LD pairs. This implies that in the case of GB, but not FCE, the two sources of negative examples provide orthogonal information that are each of their own particular use during the training process.

The specific character of the generated negative examples may explain this performance disparity. Figure 2 shows the average LD from the target variant tokens for each of the negative generation strategies. Random generation, the best performing strategy over FCE, produces negative samples on average about 8 LD from the target variant token, no matter the number generated. Logically, the mixed strategy, which performed best over GB, produced a set of samples with average an LD falling between the uniformly high LD of the randomly generated samples and the low LD of the samples generated by the LD process, which, for GB range from just below to just over 3 LD on average.

In short, FCE trained models benefit most from uniformly high LD negative examples, while GB trained models benefit most from a mixture of high and low LD negative examples. This may speak to the distinct nature of the positive examples found in these corpora. The FCE corpus is comprised of samples produced by multiple authors. However, the range of possible orthovariant forms they employ is limited by their shared intent to adhere to a standard form of English orthography as best as they can. This overriding principle could lead FCE’s variant forms to conform more closely to a centralized set of possible edits, typified by common character substitutions or phonetic misspellings – it would be understandable for a writer making a good faith attempt at producing standard English orthography to replace a "c" with a "k", but never, say, an elision apostrophe ("’"). If this is the case, much of the information the model would need to distinguish between low LD Brown candidate tokens is already contained in the fairly uniform set of possibilities demonstrated by the positive examples – the types of transformations embodied by these examples closely resembles the

set of transformations resident in the FCE corpus as a whole.

In contrast, the GB corpus contains samples drawn from multiple authors who each employ their own looser set of orthographic constraints. These authors do not attempt in good faith to adhere to a particular standard orthography. Rather, they use orthography as an expressive tool, and may not rely as heavily on further orthographic principles. Consequently, the positive examples may lose some significant amount of explanatory value. Examples of this effect drawn from the corpus can be found in Table 3.

Each variant is a phonetic or pseudo-phonetic rendering of a given word in a form of particularly motivated variant English orthography, yet each set of character-level substitutions varies to a large degree.⁴ Indeed, even though all of these forms are relatively low LD from their standard token, the set of transformation principles encoded in one teaches us relatively little about the set found in any of the others. This could explain the the mixed strategy’s superior performance on GB, as the positive examples under-determine the space of likely transformations among low LD candidates, leaving the generated low LD negative examples more room to provide useful information.

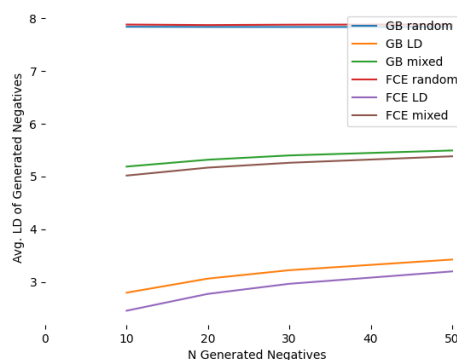


Figure 2: Average LD of Generated Negatives

5 Future Work

The complexities inherent in literary variant orthography offers many axes on which to continue these studies. Further experiments could be performed to validate the hypothesis concerning the mixed strategy’s success on the GB corpus. This could be

⁴It should be noted and acknowledged that many of these examples are due to the proliferation of explicitly racist depictions of African-Americans and other minority groups in this period of literature

Standard	Variants					
afraid	afear'd	avraid	'feerd	'fraid	'afeared	ofraid
children	childens	child'n	chillunses	chilther		
master	mars'	mars'r	'marse	mauster		
convenient	convanient	conwenient	conuenenient			
office	awffice	offfis	ohfice			
calculate	calkylate	calkelate	ca'culate			

Table 3: Samples of paired standard and variant wordforms found in GB

accomplished a number of ways, including training an additional set of models on a dataset of ortho-variants generated by other means (for example the Zéro character level adversarial benchmark (Eger and Benz, 2020)) and evaluating the performance of the negative generation strategies in the context of that dataset’s own orthographic precepts.

Additionally, future work could leverage the sentence-level contextual information included in the GB corpus to aid in string pair ranking. This could be an especially fruitful solution given the multiply-systematic nature of literary orthographic variation, as local information about the semantics of the source variant and the nature of the orthographic choices made in nearby neighbor tokens could aid in adjudicating between source-candidate pairs granted relatively similar probabilities by the neural edit distance model.

References

- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Steffen Eger and Yannik Benz. 2020. From hero to z\`eroe: A benchmark of low-level adversarial attacks. *arXiv preprint arXiv:2010.05648*.
- W Nelson Francis and Henry Kucera. 1964. A standard corpus of present-day edited american english, for use with digital computers. *Brown University, Providence, 2*.
- Sumner Ives. 1971. A theory of literary dialect. *A various language: Perspectives on American dialects*, pages 145–177.
- Gavin Jones. 1999. *Strange talk: The politics of dialect literature in Gilded Age America*. Univ of California Press.
- Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.
- Jindřich Libovický and Alexander Fraser. 2022. [Neural string edit distance](#). In *Proceedings of the Sixth Workshop on Structured Prediction for NLP*, pages 52–66, Dublin, Ireland. Association for Computational Linguistics.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- Diane Nicholls. 2003. The cambridge learner corpus: Error coding and analysis for lexicography and elt. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581. Cambridge University Press Cambridge.
- Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. [Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics?](#) In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*:

Human Language Technologies, Volume 2 (Short Papers), pages 393–400, New Orleans, Louisiana. Association for Computational Linguistics.

Eric Sven Ristad and Peter N Yianilos. 1998. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532.

Claude E Shannon. 1951. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. [A new dataset and method for automatically grading ESOL texts](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

A Hyper-parameters and model details

The hyperparameters we employ closely follow those found in (Libovický and Fraser, 2022). The RNN embedding model employs gated recurrent units (GRU) (Cho et al., 2014). The model was trained using three equally weighted loss functions: EM, binary cross entropy, and non-matching negative log-likelihood.

Name	Value
Embedding model	RNN
Embedding size	256
Hidden layers	2
Batch size	512
Validation frequency	50
Patience	10