Candidate Ranking for Maintenance of an Online Dictionary

Claire Broad[1,2], Helen Langone[1], David Guy Brizan[2]
[1]Dictionary.com
555 12th Street, Oakland, CA 94607
[2]University of San Francisco
101 Howard Street, San Francisco, CA 94105
{Claire.Broad, Helen.Langone}@dictionary.com, dgbrizan@usfca.edu

Abstract

Traditionally, the process whereby a lexicographer identifies a lexical item to add to a dictionary – a database of lexical items – has been time-consuming and subjective. In the modern age of online dictionaries, all queries for lexical entries not currently in the database are indistinguishable from a larger list of misspellings, meaning that potential new or trending entries can get lost easily. In this project, we develop a system that uses machine learning techniques to assign these "misspells" a probability of being a novel or missing entry, incorporating signals from orthography, usage by trusted online sources, and dictionary query patterns.

Keywords: Neologisms, machine-readable dictionary, ranking

1. Introduction

Dictionaries are databases in which the primary entities are words. Like a (non-temporal) database, a dictionary's contents are frozen in time (Guthrie et al., 1996; Labov, 2011; Curzan, 2012). Therefore, for a dictionary to remain relevant, new lexical entries for entirely new words – neologisms – must be added. This maintenance is important for machine-readable dictionaries (MRDs) as well as those built for human consumption.

We have created and continue to maintain a descriptive, general-purpose dictionary of American English[1]. Through our publicly-available web site, this lexical database is searchable by any of our users. Each lexical item may include one or more spellings, parts of speech, definitions, pronunciations, origins, examples of usage, and other information. Each month, our site hosts approximately 70 million users who collectively generate more than 450 million searches.

The content of our site is maintained by lexicographers. Some of this maintenance involves researching candidate lexical entries drawn from a number of sources, including unmatched queries: users' searches on our site which fail to match an item in the database. Given the size of our database and the number of unmatched queries, the work of prioritizing the items to be considered for inclusion is labor-intensive and somewhat subjective. Figure 1 contains an overview of our procedure for maintaining our dictionary.

In the interest of establishing a reasonable scope for this project, we limit our focus to single-word items. However, we believe that the same process could be effectively applied to multiword queries with some adjustment for the specific lexical considerations of phrases.

The goal of this effort is the production of a ranking for the unmatched query list to help our lexicographers identify potential candidate entries and focus attention on the items most qualified to be included in the lexicon. This work is made difficult because our candidates have no contextual information which has proven useful in the past (Weischedel et al., 1993; Nakagawa et al., 2001). This classification has the additional complexity of time sensitivity: for example, a previously rejected item may be included at a later point because of changes in usage or register (Curzan, 2012).

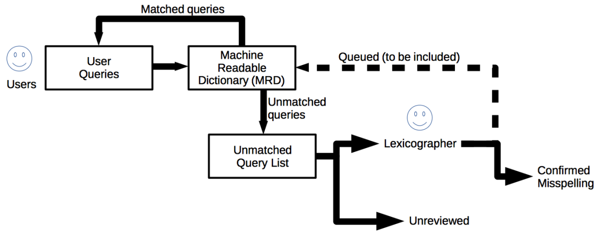

Figure 1: *Overview of Maintenance Procedure.*

While our efforts are narrowly focused on our site's content, we believe that the principles we apply here are germane to research into or practice of maintenance work for any lexical database. Since the lexicon is central to many efforts in natural language processing (Varathan et al., 2010), we see the possibility of broad appeal and obvious extensions to higher-level NLP functions, such as parsing or semantics (Al-Shalabi and Kanaan, 2004).

2. Related Work

According to one analysis (Baayenab et al., 2015) of recent and historical corpora of American English, (Davies, 2010; Davies, 2008), there may be more than 300 neologisms in the English language annually. Two sources of neologisms are *netspeak*, special "insider" language adopted by the technologically inclined starting in the 1990s, and the largely vilified *chatspeak*, which includes respellings of common words ("gr8" for "great") and abbreviations for common phrases ("brb" for "be right back") used to save time (Squires, 2010).

[1] Many non-American English lexemes (eg. "colour") are also included in our dictionary as variant spellings of their American counterparts.

Our work is related to dictionary construction in general. Although translation dictionaries were popular before his publication, Robert Cawdrey (Cawdry, 1966) is credited as having built the first monolingual dictionary of the English language in 1604 in response to the variant spellings his contemporary compatriots used, some due to the encroachment of foreign words.

The database approach for dictionaries followed more than 300 years after Cawdrey. While its stated purpose is to find historical antecedents to current language, the "Dictionary on Computer" project (Wang, 1969) describes a system for encoding lexical entries in a Chinese dictionary, alluding to the maintenance and extension of the lexicon. This work is largely steeped in the minutiae of the period in which it was written – punch cards, etc. Still, it addresses the contemporary problem of maintaining a set of lexical entries, including symbols (correlated to spelling in English) and pronunciation.

MRD-usable extensions to the database format include automatic inference of part-of-speech categories, inclusion of subcategorization frames (Boguraev et al., 1987; Sennrich and Kunz, 2014), applications to specific domains (Ji et al., 2007), or dictionaries for machine translation (Melamed, 1998; Chen et al., 1999). Building on these better-developed data are tools such as WordNet (Miller, 1995) and others which construct a graph network on the lexicon.

Theoretically, our work is inspired by the observations of (Hodges, 1972). Specifically, we do not accept the 13th century description of English spelling as chaotic. Instead, we see it as an "incompletely systematic" representation of a phonetic system which has resisted change while the spoken form of the language has welcomed loanwords and has been more inclusive of differing pronunciations, which could be anticipated (Hills and Adelman, 2015; Bromham et al., 2015; Steels and Kaplan, 1998; Longobardi et al., 2015) from the size, diversity and density of speakers from unique linguistic subcultures, dialects and registers. While English spelling is not as regular as, for example, Arabic (Al-Shalabi and Kanaan, 2004)[2] or Spanish, there are a number of patterns which we may exploit with techniques similar to those used in the phonotactic approaches for spoken language classification (Zissman and others, 1996).

3. Data

Our data is derived from two primary sources: lexical entries and queries to the site from February, 2017. The lexical entries are word types in American English which have been identified by lexicographers as being in common usage. The set of queries to the site contains all query strings (*matched* – those queries which have a corresponding entry or variant spelling in our MRD – and *unmatched*, which have no entry), as well as the monthly query count for each. The set of unmatched queries is ranked by the number of requests for each item. Of the unmatched queries, lexicographers have classified the 10000 most popular items, and the majority are verified as misspellings. We use portions of the lexical entries and queries to construct our test and training sets.

Type	True	Misspelling
Vowel Replacement	separate perceive	seperate percieve
Consonant Replacement	accommodate cynicism	accomodate synicism
Silent Letter Omission	government acquire	goverment aquire
Phonetic Spelling	rapport environment	repore enviorment

Table 1: Examples of misspell types

Our goal is the re-ranking of the unmatched queries, so that those that are strong candidates to be selected for inclusion in the dictionary appear at the top.

3.1. Test Set

The test data consists of unmatched queries from February 2017, ordered by the number of times each item was queried. Within this set, there are a few broad categories of orthographically similar character changes, primarily consisting of vowel replacements, consonant replacements, doubling or omission of characters, and silent letter omissions. We also find fully phonetic spellings in cases in which a word is pronounced differently from how it is spelled. Finally, although we find few top queries to the site involving instances of "slip of the finger" typos, we anticipate these errors. Table 1 contains examples of each of these categories.

From our observations of the historical user query patterns, we anticipate on the order of 1 valid class item for every 100 test items.

3.2. Training Set

In experimenting with different training set construction strategies, we found that downsampling the class of valid lexical entries to achieve the observed 1:100 class distribution was neither optimal nor robust – due to high variability in the valid class, the classification of the test set tended to vary greatly between trials. To mitigate this, we developed a novel ensembling technique, with the intention of increasing the influence of the 'best' valid items.

3.2.1. Valid class

We only consider single-word items, excluding all other entries. This results in over 100000 items. We prioritize the most recently added as explained below.

3.2.2. Invalid class

The only validated set of items in the invalid class consists of the previous month's 10000 most common misspellings, many of which must be rejected from use in the training set because they were also commonly searched in the current month and are therefore present in the test set. Because of this overlap, we have fewer examples (7871) of misspellings than required to match the distribution of the test

[2] The case for Arabic spelling may be more complicated than we make it out to be. For example, Modern Standard Arabic–a common "second dialect" of many in the Arabic-speaking world–may have many of the issues we allude to in English.

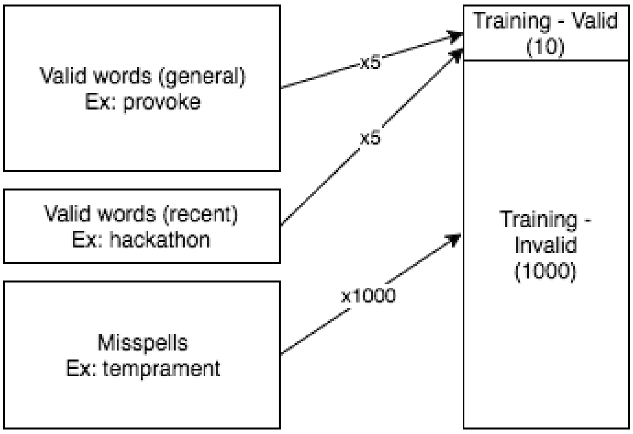

Figure 2: *Construction of a training set.*

set. We supplement the invalid class by generating the following random strings and ensure that there are no matches to the variant set.

- Plausible misspellings: a random lexical entry is selected and a string is replaced by an orthographically similar one, ideally replicating the types shown in Table 1
- Typos: a random item is selected from the primary lexicon and a character is replaced with one of its neighbors on the keyboard
- Random character strings of length up to 20 characters

4. Methods

Using the data described above, we generate a hypothesized probability that a candidate item from the test set will be adopted as a new lexical entry. The test set items are reordered by probability to create the final ranking.

4.1. Feature Extraction

The model includes a variety of features extracted from the attributes of the item, the query data, and word usage on the Internet at large.

4.1.1. Orthographic Features

We extracted the following orthographic features for each item in our training and test sets:

1. Presence of a known prefix (suffix) at the beginning (end) of the word
2. If (1) is true, is the remainder of the item a word from the variant set?
3. Presence of Greek/Latin roots (determined by whether any item in predetermined set of possible roots is present as a substring)
4. Count and proportion of vowels/consonants
5. Character length of the lexical item
6. Portmanteau: is the item a concatenation of two variant items – that is, does any single split of the item result in two substrings that are in the variant set

In addition, we encode all pairs of characters (including '-' and the starts and ends of words) using a variant of term frequency-inverse document frequency (tf-idf), where the frequency of a character pair in the item is compared against the frequency of that character pair in the full set of training items.

4.1.2. Query Pattern Features

To the orthographic features, we add data on the items' popularity. The goal with the query pattern features is to isolate not only the item's current popularity, but also whether it is of recent interest. A true neologism may have a large number of queries concentrated in a recent period of time, whereas very common misspellings would have consistently high queries. Therefore, we also include two percent change features:

- Over mean: comparing the number of queries of the item this month against its mean monthly queries in past months
- Year-over-year: comparing this month's queries against the queries of this item in this month of last year (to account for items of seasonal interest)

Furthermore, we include the percentage of months for which we have volume for user query data that the item.
For scalability reasons, these calculations are not performed on items that do not have queries in the current month. The traffic features for these items are populated with placeholder values within class that were determined via iteration to be most conducive to the significance of these features and result in a more effective model.

4.1.3. Usage in the Wild

Finally, we consider whether an item is being used elsewhere on the Internet. We found that one ideal source for this information was the Twitter feeds of news organizations, as they can be expected to be quite rigorous with spelling and include a variety of trending terms. We used the Twitter API to extract the language usage of 100 Twitter feeds from broadcast and print media outlets such as ABC News, Yahoo News and the Huffington Post.
273 of the 8115 items in the test set are present in the Twitter corpus. Some misspellings are present, along with a wide variety of strong candidate keywords. For example:

- Typos/misspellings: aquired, seige, beacuse
- Proper nouns: Supercell, Starbucks, Chromebook
- Coinages: deflategate, yuge, cuck
- Slang: turnt, rekt, janky
- Neologisms: petrichor, misophonia
- Loanwords: queso, agua, deux

4.2. Prediction Model

We generate a number of classifiers, each trained on a very small portion of the valid items, and prioritize those which are most effective at classifying misspellings correctly. For 1000 iterations, a small training set is randomly sampled,

Type	Count	Query Rank	Model Rank
Confirmed Misspelling	4304	3257	4787
In the Queue	35	3689	2558
Neither/unvetted	3774	4876	3240

Table 2: Change in average ranking by subset

consisting of 10 valid items (half recent additions, half from the main set of lexical entries) and 1000 invalid items, as shown in Figure 2. Using the random forest implementation in scikit-learn (Pedregosa et al., 2011), a forest of 10 estimators is trained on each set and persisted for use against the test set. We also store the mean score of this forest on predicting the other training sets. (For ease of computation, we use a random sample of 5 sets.) Each of these forests is then used to predict the probability of validity for the items in the test set, and their predictions are ensembled using the forests' respective scores as weights.

5. Results

Tested against query data from February 2017, we see strong results in predicting potentially valid items. We assess these results with two metrics.

Our first metric is the mean rank of the items in each class within the list of confirmed misspellings. With the caveat that an item may be accepted at a later date, we use this to assess, in a general sense, the rankings of misspellings among the predictions. Our assumption is that no more than one or two such shifts to validity would occur at a time, and thus would not skew this figure too egregiously.

Our second metric employs the "queue," items identified by the lexicographers as valid but not yet included in the dictionary for procedural reasons.

The test set consists of 8115 items:

- 37 items in the queue
- 4306 items on the confirmed misspells list

There are two items which are in both categories – instances of the aforementioned edge case where an item previously considered a common misspelling has subsequently been accepted as a valid item for our MRD. Table 2 shows the average ranking of test set items that are within one or neither of these sets, comparing rank as determined by query count alone versus the rank generated by the model.

Selected test items, with query count ranking and predicted validity ranking in parentheses:

- Slang + loanwords: boujee (45; 1), hola (32; 7), hygge (445; 9), adulting (5656; 19)
- Pop culture: pikachu (132; 4), harambe (842; 5), moana (1994; 8), festivus (4445; 33)
- Tech: youtuber (1476; 6), blockchain (7202; 12), ransomware (6258; 52)
- Politics: CPAC (4335; 3), alt-right (3826; 14), post-truth (3411; 16)

6. Discussion

The average ranking for all items in the queue, 2707, is well above the midpoint of 4058, as well as an improvement from the average ranking based on query count alone of 3257. Furthermore, the items which are on neither list have a higher average ranking than the confirmed misspells (3240 vs 4788). This is positive, as this is in many ways our target group – items which are neither obvious misspellings, nor obvious candidates, and thus have not been identified previously. Finally, the confirmed misspellings have moved down in rank, from 3257 to 4787.

We believe that these results indicate that our system is capable of identifying the same candidate terms that would be chosen by a lexicographer, as well as additional terms that would otherwise have stayed buried in the "misspell" bucket. Indeed, the rankings produced by this system have been adopted by our lexicographical content team as a tool for identifying keywords that merit further research.

We also believe that these results show the merit of our orthographic approach as a surrogate for American English pronunciation. We see evidence of this in that adopted items which are Anglicized, such as "boujee" are ranked higher than un-modified loanwords like "hygge." We anticipate applying this approach to different data and domains.

7. Conclusions and Future Work

The approach we describe is limited to single tokens queried on the site because of our focus and our available data. We are planning to verify our results with data from external sources and investigate how the same dictionary construction could be automated to benefit NLP applications which have temporal considerations.

Although this effort was naturally limited to American English by virtue of our data source and use case, we believe that the same basic principles could be applied to many languages. While many languages do not have the same complexities of orthography, they may be influenced by external pressures, so maybe more subject to the adoptions of calques. Therefore, our approach may require adjustments to the orthographic features to best suit the given language.

For future work, we will extend our approach to include multi-word expressions. To that end, we are also interested in applying new approaches such as deep learning using character-level embeddings on an LSTM network.

8. Bibliographical References

Al-Shalabi, R. and Kanaan, G. (2004). Constructing an automatic lexicon for arabic language. *International Journal of Computing & Information Sciences*, 2(2):114–128.

Baayenab, H., Tomascheka, F., Gahlc, S., Ramscara, M., and Baayen, R. H. (2015). The ecclesiastes principle in language change.

Boguraev, B., Briscoe, T., Carroll, J., Carter, D., and Grover, C. (1987). The derivation of a grammatically indexed lexicon from the longman dictionary of contemporary english. In *Proceedings of the 25th annual meeting on Association for Computational Linguistics*, pages 193–200. Association for Computational Linguistics.

Bromham, L., Hua, X., Fitzpatrick, T. G., and Greenhill, S. J. (2015). Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences*, 112(7):2097–2102.

Cawdry, R. (1966). *Table Alphabetical of Hard Usual English Words*. Scholars Facsimilies & Reprint.

Chen, A., Kishida, K., Jiang, H., Liang, Q., and Gey, F. C. (1999). Automatic construction of a japanese-english lexicon and its application in cross-language information retrieval. In *In Joint ACM DL/ACM SIGIR Workshop on Multilingual Information Discovery and AccesS (MIDAS*. Citeseer.

Curzan, A. (2012). *The Secret Life of Words: English Words and Their Origins*. The Great Courses.

Davies, M. (2008). The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, 25(4):447–464.

Davies, M. (2010). The corpus of historical american english: 400 million words, 1810–2009,(2010). *URL: http://corpus.byu.edu/coha/*.

Guthrie, L., Pustejovsky, J., Wilks, Y., and Slator, B. M. (1996). The role of lexicons in natural language processing. *Communications of the ACM*, 39(1):63–72.

Hills, T. T. and Adelman, J. S. (2015). Recent evolution of learnability in american english from 1800 to 2000. *Cognition*, 143:87–92.

Hodges, R. E. (1972). Theoretical frameworks of english orthography. *Elementary English*, 49(7):1089–1105.

Ji, L., Lu, Q., Li, W., and Chen, Y. (2007). Automatic construction of a core lexicon for specific domain. In *Advanced Language Processing and Web Information Technology, 2007. ALPIT 2007. Sixth International Conference on*, pages 183–188. IEEE.

Labov, W. (2011). *Principles of linguistic change, cognitive and cultural factors*, volume 3. John Wiley & Sons.

Longobardi, G., Ghirotto, S., Guardiano, C., Tassi, F., Benazzo, A., Ceolin, A., and Barbujani, G. (2015). Across language families: Genome diversity mirrors linguistic variation within europe. *American journal of physical anthropology*, 157(4):630–640.

Melamed, I. D. (1998). Empirical methods for mt lexicon development. In *Conference of the Association for Machine Translation in the Americas*, pages 18–30. Springer.

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

Nakagawa, T., Kudo, T., and Matsumoto, Y. (2001). Unknown word guessing and part-of-speech tagging using support vector machines. In *NLPRS*, pages 325–331. Citeseer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Sennrich, R. and Kunz, B. (2014). Zmorge: A german morphological lexicon extracted from wiktionary. In *LREC*, pages 1063–1067. Citeseer.

Squires, L. (2010). Enregistering internet language. *Language in Society*, 39(04):457–492.

Steels, L. and Kaplan, F. (1998). Spontaneous lexicon change. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 1243–1250. Association for Computational Linguistics.

Varathan, K. D., Sembok, T. M. T., and Kadir, R. A. (2010). Automatic lexicon generator. In *Information Retrieval & Knowledge Management,(CAMP), 2010 International Conference on*, pages 24–27. IEEE.

Wang, W. S. (1969). Project doc: Its methodological basis. In *Proceedings of the 1969 conference on Computational linguistics*, pages 1–22. Association for Computational Linguistics.

Weischedel, R., Schwartz, R., Palmucci, J., Meteer, M., and Ramshaw, L. (1993). Coping with ambiguity and unknown words through probabilistic models. *Computational linguistics*, 19(2):361–382.

Zissman, M. A. et al. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on speech and audio processing*, 4(1):31.