The Etymological Machine

Anonymous ACL submission

Abstract

The purpose of this paper is to build a machine to discover the proportions of word usage according to etymological origins. As a short term goal, we intend to produce a user application for linguistic outreach by allowing people to play with the concept of etymology, as well as clearing up a misconception spreading around the Internet. This paper concerns the process of constructing a database for this purpose, including the problem presented by classifying entries in an etymological dictionary into simple categories based on language of origin. We obtain up to 82% accuracy with bag-of-Words, character-incidence, date-extraction, and syllables present in the word itself as features for SVM classification.

1 Introduction

The inspiration for this paper is an image that is occasionally shared around the Internet originating from Wikipedia that claims to show the relative proportions of word origins in English, and it shows that over 50% of English vocabulary is Latinate. This is then used as evidence that English is, in fact, a Romance Language. The argument often paired with it goes something like, If modern linguists were to look at English as it is, without the benefit of history, they would see all of the Latin and French words and conclude that English is Romance.

Naturally, this is fundamentally incorrect (Bammesberger, 1992), but saying so is hardly effective outreach. In this paper, we move to demonstrate that the above chart is misleading because the chunks labeled "Latin" and "French" contain words that are extremely rare in actual usage, like

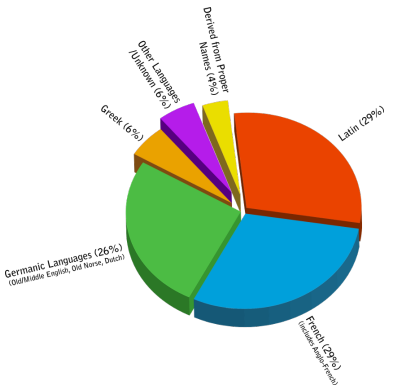

Figure 1: Misleading chart from the internet (Wikipedia user Murraytheb, Public Domain)

talaric, *lexiphanicism*, and *calcography*. An actual analysis of a text would show a very different proportion. More formally:

- A full-dictionary analysis of the origins of words in English will return a chart like the above.

- Typical English texts will be mostly Germanic, upwards of 60%

- One exception will be scientific (especially medical) texts, which should see a raised proportion of Latin and Greek words

- Another exception will be legal texts, which will contain a high proportion of French words (because most of the English legal vocabulary was set during the reign of Henry I, who spoke French).

Before doing this analysis, there needs to be a database of English words categorized into their source language. Unfortunately, we could find no database of words classified into etymological groups, so we needed to make one from one of many human-readable dictionaries that do exist and are freely available on-line. This is a text classification task, one which will be accomplished with both pattern-matching and machine-learning methods.

2 Data Preparation and Pattern-matching Classification

2.1 Source

We crawled the Online Etymological Dictionary (Douglas Harper, 2016) to collect a dictionary that maps words of English to their etymological entries (which we are going to call **definitions** from now on for convenience).

2.2 Class selection

The number of words from each source language follows a Zipfian distribution, with a small number of source languages covering 80-95% of total tokens, with the rest distributed across a large number of categories (Finkenstaedt and Wolff, 1973; Williams, 1986). The largest lexical donors to modern English, and also the categories listed in the image above, are Old English, French, Old Norse, Latin, and Greek. All the languages but these biggest few will be lumped in an "Other" category. The classes therefore will be as follows: Old English, French, Old Norse, Latin, Greek, and Other.

2.3 Hierarchy

Some of the above languages have a relationship to each other. This means that many of the entries in the dictionary will have followed etymological "chains". For example:

> polysemy - 1900, from French polysémie, from Medieval Latin polysemus, from Greek polysemos, from poly- + sema

It is necessary to build these hierarchical paths into the classifier, to make sure that, should they appear in non-chronological order in the dictionary, that we can still get the right answer. This is not very difficult, as, barring much cross-interference, we see the following chains

- French > Middle French > Old French > Medieval Latin > Latin > Greek > Proto-Indo-European (PIE)
- Greek(modern, koine, classical) > PIE
- Old English > Ingvaeonic (Old Frisian, Old Saxon) > Proto-West-Germanic > Proto-Germanic > PIE
- Old Norse > Proto-North-Germanic > Proto-Germanic > PIE

This is not to say that our model will have all these categories, only that terms are likely to follow these paths, and so they can be used as covariant features in training. Also, do not read the above to imply a genetic relationship, these are simply the most likely paths that lexemes will follow on the way to English.

2.4 Deterministic Classification and Result

As a preliminary to the statistical classification to obtain a full dictionary, it was necessary to classify as many dictionary entries by deterministic means as possible. This was done with pattern matching, simply searching for the terms above. After resolving an issue with definitions that had no content but merely linked to other words (resolved by recursively copying the classification of the link) we arrived with the class cardinality (See Table 1).

Language	Count	%
English	11300	24.71
French	12712	27.80
Norse	495	1.08
Latin	6946	15.19
Greek	1465	3.20
Other	11131	24.34
Total	44049	100.00

Table 1: Pattern-match classification results, counts and percentages

There were an additional 1674 words that were not categorizable with this method. This is due to an unidentified bug in the code that somehow allowed these words to be added to the table with an empty-string category. There are not any immediately apparent similarities between these words or dictionary entries, varying from *cumulonimbus* to *skunk*, so these will be set aside during training, and categorized using the most promising model to generate a final product.

A manual check of the above categorizations from a random sample of 200 items produced an accuracy rating of 82%. There were some items that received no classification for reasons unknown, these will be classified using the machine learning classifier described in the next section.

3 Statistical Method and Feature Selection

After testing several model kernels on a subset of the data, using various feature sets, it was found that a linear-kernel SVM outperformed the others, so this is what was used to test the following features

3.1 Features

As a text classification problem, the natural feature to include is bag-of-words. This particular model does not have the typical problem with bag-of-words of *hapax legomena*, because we have the full dataset and all of the possible words from the beginning, so there can be a feature dimension for every word, even if it doesn't happen to be in the training set (Boulis and Ostendorf, 2005).

For the purposes of exploring classification effects, we used a list of stop-words that included the class labels themselves. By not allowing the word "English" definition to be shown to the SVM, we hoped to see whether the machine could find other words that distinguish themselves as features of the Old English category besides "English" itself, and likewise for the other classes.

We hypothesize that it may be possible to classify words based on features present in the word itself, and that the information present in the word itself may be as good of a predictor as bag-of-words. We separated each word into syllables and treated the syllables-present as a feature vector. We use the CELEX lexical databases of English[1] to train a model for determining the syllable boundaries.

Thirdly, we propose as a new feature, "characters present". This is a vector of binary values, each of which corresponds to a character in the training set. Our hope in proposing this feature is that the model will learn that characters like é and ç are more likely to be present in the definitions of French-origin words, and ð and þ in the definitions of Old English-origin words.

The last feature we attempted is "first attested century", extracted from the first two digits of the first three or four consecutive digits in the definition, which is usually the first century in which a word appears in English. The hope is that the classifier will learn that Old English words are usually attested before the 11th century, French words start appearing in the 12th-15th centuries, etc. Since not all of the definitions have this feature, it is tested only as an addition to the others.

[1] https://catalog.ldc.upenn.edu/ldc96l14

4 Experiment and Results

4.1 Statistical Classification

The following table shows the 5-fold cross-validation results of a classifier trained using the above features on the Etymonline data set. See Table 2.

Features	Accuracy
Syllables	0.48
BagofWords	0.71
Characters	0.72
Syllables + BagofWords	0.73
Syllables + Characters	0.75
BagofWords + Characters	0.81
BagofWords + Characters + Syllables	0.82
Syllables + Characters + Century	0.75
BagofWords + Characters + Century	0.81
All Four Features	0.82

Table 2: 5-fold cross-validation test results on combinations of features against a random 40% sample of the full dataset

The century feature is not as effective as expected. In fact, it seems to have little to no effect on the results at all, each of the tests including the century feature being indistinguishable to two significant digits from the associated test without that feature.

We then took the best performing classifier: that with all four features, and used it to classify the erroneous words that were missing from the initial classification to generate a final dictionary (See Table 3).

Language	Count	Percent
Old English	11751	25.7
French	12848	28.1
Norse	504	1.1
Latin	6949	15.2
Greek	1463	3.20
Other	12208	26.7
Total	45723	

Table 3: Proportions in statistically categorized lexicon classified using all four features, including the errors missing from the previous test

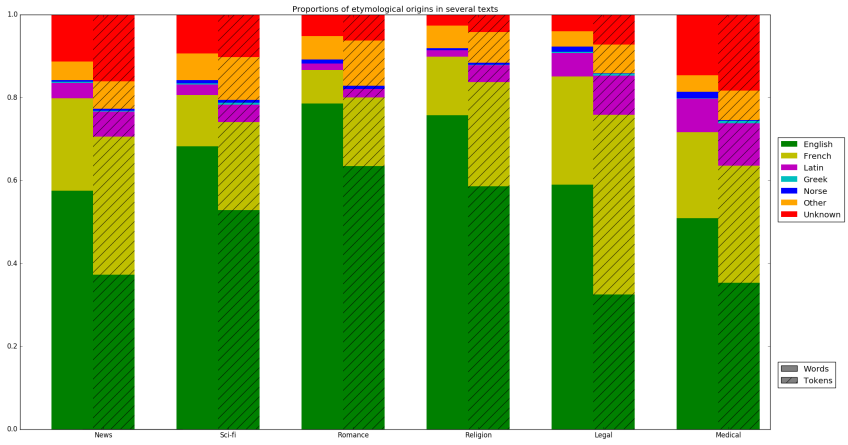

Figure 2: Etymological Analysis of six Brown texts of roughly equal length. Comparison of absolute word counts and unique tokens. Top to bottom: Unknown, Other, Norse, Greek, Latin, French, Old English

4.2 Real-text Tests

Since the etymological entries in Etymonline are labeled with Part of Speech, and part of the demonstration that we hope to present relies on differences between works of different genres, we selected the Brown corpus for testing, since it is structured by genre and tagged (See Figure 2).

Texts used:

- Philadelphia Inquirer 1961 Politics Section (ca09), 2234 words

- Stranger in a Strange Land, Robert Henlein (cm01), 2486 words

- The Snake, Evrin D Krause (cp26), 2520 words

- Faith Amid Fear, Peter Eversveld (cd07), 2359 words

- Public Laws of the 87th Congress (ch09), 2451 words

- Wound-Tumor Virus Antigen, Nagaraj and Black (cj16), 2258 words

5 Related Work

The only similar project we have found is a post on the Ideas Illustrated blog by Mike Kinde (Kinde, 2012). Kinde's project was similar in kind (perhaps without the explicit goal of refuting a somewhat common misconception), but not in practice, as he used only four small corpora of one paragraph each, and tagged each word manually.

Part of the inspiration for the Etymachine project is to replicate the visualization results Kinde produced automatically, so that users can feed in their own texts and see for themselves how much Romance is in their speech.

6 Conclusion and Future Perspectives

The word-syllables feature underperformed our expectations. The classifier beat random chance, but the effect was less pronounced than expected. We have plans for future work to apply deep-learning techniques to this problem to attempt to classify words into etymological categories without relying on the dictionary at all.

Our hypotheses concerning analysis of English texts seems to be confirmed. The more conversational the tone of a text, the further it goes above 60% Old English, with the Romance novel, "The Snake" having the least Romance of all in both the word and token counts, and the medical and legal texts having the most.

We plan to use this work as the basis for a user application that can be shared with the public to clear up the misconception that English is heavily Romance in vocabulary, and to inspire interest in historical and computational linguistics.

We propose that the proportional values returned by the Etymology Machine can be used as a feature for text classification. There is a correlation, for example, between legal texts and a large number of French lemmas. This could be used as an additional feature for genre classification. We are planning on further research on larger corpora.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Alfred Bammesberger. 1992. Chapter 2: The Place of English in Germanic and Indo-European. In *The Cambridge History of the English Language*, Cambridge University Press, chapter 2, pages 26–66.

Steven Bird, Edward Loper, and Ewan Klein. 2016. Natural Language Toolkit (NLTK). http://www.nltk.org/.

C Boulis and M Ostendorf. 2005. Text Classification by augmenting the bag-of-words representation with reduncancy-compensated bigrams. *SIAM International Conference on Data Mining* .

Douglas Harper. 2016. Online Etymological Dictionary. http://www.etymonline.com/.

Thomas Finkenstaedt and Dieter Wolff. 1973. *Ordered profusion: Studies in Dictionaries and the English Lexicon*. C. Winter.

W N Francis and H Cucera. 1979. A Standard Corpus of Present-Day Edited American English, for use with Digital Computers.

Mike Kinde. 2012. Visualizing Word Origins.

F Pedregosa, G Varoquaux, A Gramfort, V. Michel, B Thirion, O Grisel, M Blondel, P. Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011a. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

F Pedregosa, G Varoquaux, A Gramfort, V. Michel, B Thirion, O Grisel, M Blondel, P. Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011b. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Joseph M. Williams. 1986. *Origins of the English Language. A Social and Linguistic History*. Free Press.