UCD : Diachronic Text Classification with Character, Word, and Syntactic N-grams

Terrence Szymanski
Insight Centre for Data Analytics
School of Computer Science and Informatics
University College Dublin, Ireland
terrence.szymanski@ucd.ie

Gerard Lynch
Centre for Applied Data Analytics Research
University College Dublin, Ireland
gerard.lynch.@ucd.ie

Abstract

We present our submission to SemEval-2015 Task 7: Diachronic Text Evaluation, in which we approach the task of assigning a date to a text as a multi-class classification problem. We extract n-gram features from the text at the letter, word, and syntactic level, and use these to train a classifier on date-labeled training data. We also incorporate date probabilities of syntactic features as estimated from a very large external corpus of books. Our system achieved the highest performance of all systems on subtask 2: identifying texts by specific time language use.

1 Introduction

This paper describes our submission to the SemEval-2015 Task 7, "Diachronic Text Evaluation" (Popescu and Strapparava, 2015). The aim of this shared task is to evaluate approaches toward diachronic text analysis of a corpus of English-language news articles from The Spectator[1] archive, originally published between 1700 and 2014.

We solely address subtask 2: "texts with specific time language usage." The goal of this subtask is to infer the composition date of a text based on implicit clues in language of the text, as opposed to overt mentions of datable named entities or events. This task has inherent utility, for example, for historians dating texts in an archive with no external datable properties. However, it is equally interesting as an investigation into methods for quantifying changes in language and writing style over a period of centuries.

We approach this task in a similar manner as previous work on stylistic text classification (Argamon-Engelson et al., 1998) in that we aim to model stylistic, rather than topical, features of the text. From each text we extract a variety of character, lexical, and syntactic features, as described in section 3. We also use a set of syntactic features whose frequencies over time have been estimated from a very large corpus of books (Goldberg and Orwant, 2013). While many of these features have previously been used for stylistic analysis, our approach is not to model *style* per se. Many types of variation may be captured indirectly by our features: the spelling, typography, lexicon, and grammar of English have changed markedly over the past centuries, as has the genre of news writing. We consider any time-correlated variation to be useful for dating.

2 Data

We used the two training sets of texts provided by the challenge organizers for subtask 2. After removing errors (repeated items, items containing no text, items with invalid dates), our training set consisted of 4130 items. Each item contains the text of a snippet of news, typically consisting of a few sentences (the average length of a text is 70 words), and three year-range labels: one for each of the Fine (6-year), Medium (12-year) and Coarse (20-year) granularities specified in the task.

The given labels are not well-suited for classification, since the set of labels used for one text is not necessarily the same as the set of labels used for an-

[1] http://www.spectator.co.uk/.

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 879–883,
Denver, Colorado, June 4-5, 2015.

other text. For example, here are the labels provided for two texts in the training set:

```
<text id="378rn324911597">
<textF yes="1698-1704" no="1705-1711" ...
<textM yes="1695-1707" no="1708-1720" ...
<textC yes="1691-1711" no="1712-1732" ...

<text id="74gi329732114">
<textF yes="1699-1705" no="1706-1712" ...
<textM yes="1696-1708" no="1709-1721" ...
<textC yes="1692-1712" no="1713-1733" ...
```

These two texts are very close in date, yet have completely different (and incomparable) year ranges. Therefore, we create our own non-overlapping year-range classes at 6-, 12-, 20-, and 50-year levels. We assume that the true date of a text is the midpoint of the "yes" year ranges and assign a non-overlapping class appropriately. All of our training, cross-evaluation, and prediction is done using these non-overlapping classes. To make predictions for our official submission, we predict whichever given year range has the greatest overlap with our predicted class.

The training data is unevenly distributed over the possible range of years from 1700 to 2014. Just three years (1717, 1817, and 1897) account for 11% (444 of 4130) of the training instances, while 48% (150 of 314) of the years in the possible range are unattested in the training data. Overall, there is a general bias towards earlier years in the time range. We do not attempt to control for this bias in the data, since we assume that the test data will be drawn from a similar distribution. While the uneven distribution may artificially boost the accuracy of our classifiers, the baseline classifier captures this effect.

3 Features for Classification

We extract four types of features from each text: character n-grams (`Char`), part-of-speech tag n-grams (`POS`), word n-grams (`Word`) and syntactic phrase-structure rule occurrences (`Syn`). We refer to the combined feature set as `CPWS`. (Stamou, 2008) surveys diachronic classification of literary text and finds that parts of speech, character frequencies, and function word frequencies are all used in chronologically dating text composition. Part-of-speech and word n-grams have been used for stylistic text classification (Argamon-Engelson et al., 1998), and syntactic phrase-structure rules have successfully been used as stylometric features for detecting deceptive writing in online reviews (Feng et al., 2012). We have not included document-level stylistic features (e.g. average sentence length, average word length, lexical richness, lexical density, and readability measures) although they have been used successfully for diachronic stylistic analysis (Štajner and Zampien, 2013), and could be incorporated in our classification approach. However, our n-gram features may capture features such as sentence length by proxy (e.g. in the frequency of periods).

Character n-grams are an expressive feature set which can capture variation on the morphological level (word stems), syntactic level (gaps between words and punctuation) and also word-level frequency fluctuations (prepositions and conjunctions). Character bigrams were used previously on Latin text by (Frontini et al., 2008) to date the Donation of Constantine, a study which did not verify the work as a forgery but did place it in the correct stylistically implied period.[2] Additionally, character n-grams are used in stylometric tasks such as authorship attribution (Keselj et al., 2003) and detection of *translationese* (Popescu, 2011).

All n-gram features were extracted for $n \in \{1, 2, 3\}$ using an in-house Java concordancer. Punctuation and spacing was not modified during this process, although case information was discarded. No stop words were removed. Raw frequency counts of the features were used in the process, and those features with less than 20 occurrences in the entire corpus were discarded. Texts were parsed with the Stanford parser,[3] and the 250 most-frequent syntactic rules in the training set were used as features. The dependency parse was also produced and used as described below.

3.1 Google Syntactic N-grams

As an external source of data, we used the Google Books Syntactic N-Grams (GSN) database (Goldberg and Orwant, 2013). Due to the size of the datasets and time limitations, we focused solely on the *nodes* collection of the *Eng-1M* corpus, a sample of 1 million English-language books dating from 1520 to 2008. Each data point in the *nodes* collec-

[2]Verification of forgery was based on false information contained in the text, rather than stylistic idiosyncracy.

[3]http://nlp.stanford.edu/software/lex-parser.shtml

tion is a POS-tagged word and the label of the syntactic dependency between that word and its head, which gives a sense of the word's syntactic function in a given sentence. For each node, the total number of occurrences in each year is provided.

Because the GSN database is particularly sparse for years prior to 1800, we smoothed all node counts by averaging over the five nearest years with nonzero counts. Then the smoothed counts are normalized within each year to estimate the probability of a node in a given year.

We use a Naive Bayes classifier ($\texttt{Google}_{nb}$) to predict the most likely year for a given text, represented as a set of nodes extracted from the dependency parse. We also produce a GSN feature set consisting of 308 features (one for each year in the range 1700-2008), whose values are based on the total log probability of the text in that year, normalized to the interval [0,1] for each text. The normalization controls for text length and allows comparison between texts. These features are then used in the combined `CPWS+G` classifier.

3.2 Feature Informativeness

When all features are combined, the GSN features are the most predictive. In order to assess the effectiveness of the other features and also to reduce the feature set for classification, we performed attribute selection using the Weka data mining software (Hall et al., 2009). Table 1 shows the top-ranked `CPWS` features using the the 50-year class labels, using Weka's information gain attribute evaluation with 10-fold cross-validation.

Rank	Attribute	Type	Rank	Attribute	Type
1	NN	`P-1`	10.9	t	`C-1`
2	i	`C-1`	11.7	l	`C-1`
3.5	␣	`C-1`	13.3	o	`C-1`
3.9	. → .	`S`	13.8	.	`W-1`
5	ROOT → S	`S`	15.7	[' d]	`C-2`
5.8	a	`C-1`	15.9	[␣ .]	`C-2`
7.3	e	`C-1`	16.3	r	`C-1`
8.1	.	`C-1`	18.6	[JJ NN]	`P-2`
8.5	n	`C-1`	19.3	JJ	`P-1`
10.7	s	`C-1`	19.8	c	`C-1`

Table 1: Top 20 `CPWS` features using Information Gain

The rankings show that the character n-gram features were particularly expressive in capturing temporal variation, yet it can be difficult to assign a linguistic motivation to them. Because our feature counts are not normalized by text length, many of these features may simply be redundantly capturing an overall length effect.

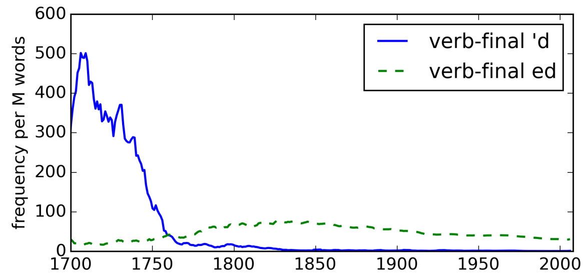

Figure 1: Changing frequencies of verb endings in the Google Books English corpus, 1700-2000.

However, some meaningful features can clearly be recognized, such as [' d], refering to the 18th century abbreviation of *-ed* as a past participle verb ending in English. The frequency of verb-final *'d* in the POS-tagged Google N-grams dataset (Lin et al., 2012), shown in Figure 1, illustrates how use of this linguistic feature has declined over time.

Another highly informative feature is the character bigram [␣ .]. In some texts, punctuation has been separated from the neighboring words with a space, possibly due to OCR errors on older texts.

4 Classification and Evaluation

We employ attribute selection as above in all of our cross-validation experiments and our official submission. Table 2 illustrates how SVM classification accuracy varies with feature set size. The value of 4000 features was chosen to maximize accuracy while minimizing running time, and was used to produce all of the results described in this paper.

$\lvert F \rvert$	6-Year	12-Year	20-Year	50-Year
4000	**37.61**	**39.30**	52.07	**67.74**
2000	35.75	37.61	50.77	67.59
1000	32.55	37.26	51.24	67.53
500	33.09	38.62	**52.22**	65.94
200	33.77	35.28	50.41	64.07
100	31.87	33.62	47.10	60.32
50	29.57	31.82	44.91	57.07

Table 2: Effect of feature set size ($\lvert F \rvert$) on classification accuracy. (`Char+POS+Google` features)

Assigning a date to a text is not a typical classification problem, because the classes are not independent of one another. We experimented with SVM regression, but this produced lower accuracy than the SVM classifier. Ordinal classification is a method that may be used when classes exhibit a natural order, as in this task. We performed some experiments with the Weka implementation of ordinal regression (Frank and Hall, 2001) using a SVM base classifier, but these produced lower accuracy than the standard SVM classifier. Therefore, we used a standard multi-class SVM classifier for all of our evaluations and predictions.

System	6-Year	12-Year	20-Year	50-Year
`Baseline`	10.4	12.6	20.5	36.6
$\texttt{Google}_{nb}$	10.9	18.7	31.7	52.4
`Char`	36.1	38.4	47.9	64.5
`POS`	24.6	26.8	36.3	53.6
`Word`	26.1	29.6	37.2	54.6
`Syn`	23.4	26.3	38.5	54.6
`CPWS`	36.9	40.1	50.7	67.8
`CPWS+G`	**41.5**	**45.9**	**55.3**	**73.3**

Table 3: Classification accuracy of various feature sets, using 10-fold cross-validiation on the training data set.

Table 3 lists the cross-validation classification accuracy for our various models. The baseline classifier looks only at the class labels and chooses the most frequent class. The $\texttt{Google}_{nb}$ classifier is a Naive Bayes classifier using only the GSN probabilities and assuming a uniform prior over years. This represents a classifier with no domain knowledge of the text genre or date range distribution.

The remaining rows show the results for SVM classifiers trained independently on each of the four stylistic feature sets. While each feature type outperforms the baseline, the character n-gram features are clearly the single most effective feature type. The combination of all four features together (`CPWS`) outperforms any single feature set individually, and this represents the maximal performance we achieve using solely the training data provided by the task organizers.

The final row shows the performance of a SVM classifier using all of our stylistic features plus features derived from the GSN probabilities. This achieves the highest accuracy and this is the system we submitted to the task.

Table 4 shows the official results of the `CPWS+G` classifier, trained on the full training set and evaluated on a test set of 1041 texts whose true dates were unknown to us. The accuracy values are in line with our cross-validation scores. The score is a weighted classification metric that rewards predictions that are not fully correct but are near the correct date. The third row lists the mean deviation of our predictions from the true date. By all three measures, our system was the top performing submission to this subtask.

	Fine (6-year)	Medium (12-year)	Coarse (20-year)
Accuracy	46.3	47.3	54.3
Score	0.7592	0.8466	0.9104
Avg. Years Off	14	19	19

Table 4: Official results on the SemEval test data.

Our 73.3% accuracy on the 50-year class may be loosely compared to (Mihalcea and Nastase, 2012), who achieve 62% classification accuracy dating words in context to 50-year epochs. Their task, word epoch disambiguation, is comparable but different: they classify words, not texts, using local context features and a targeted set of 165 words.

5 Conclusion

We have shown that a stylistic classification approach is capable of accurately predicting the date when a text from the sample category was written. Additionally, our approach is straightforward to implement and can function well using only a moderate sized sample of training data, although its accuracy can be improved by incorporating features trained from a large external corpus.

We cast a wide net in order to produce a large feature set and allow the classifier to select whichever features most improved the classification accuracy. While this produces good classification results, it remains difficult to interpret the linguistic or stylistic significance of the most-predictive features. It is also unknown how the results would differ on other data sets in different languages, genres, or time periods. In addition to the features we have explored, there are a number of others, such as sentence length, capitalization, and lexical richness measures which might be considered in future work.

Acknowledgments

This work is supported by Science Foundation Ireland through the Insight Centre for Data Analytics under grant number SFI/12/RC/2289 and Enterprise Ireland through the Centre for Applied Data Analytics Research under grant number TC 2013 0013.

Thanks to Mark Keane for his feedback and suggestions, particularly on the use of syntactic features for dating. Thanks also to the Insight Centre Future of News discussion group for their feedback on a presentation of an early version of this work.

References

Shlomo Argamon-Engelson, Moshe Koppel, and Galit Avneri. 1998. Style-based text categorization: What newspaper am I reading? Technical report, AAAI.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of ACL 2012: Short Papers*, pages 171–175.

Eibe Frank and Mark Hall. 2001. A simple approach to ordinal classification. Technical report, University of Waikato.

Francesca Frontini, Gerard Lynch, and Carl Vogel. 2008. Revisiting the 'Donation of Constantine'. In *Proceedings of AISB 2008*, pages 1–9.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of English books. In *Proceedings of *SEM 2013*, pages 241–247.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

Vlado Keselj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of PACLING 2003*, pages 255–264.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of ACL 2012*.

Octavian Popescu and Carlo Strapparava. 2015. Semeval-2015 task 7: Diachronic text evaluation. In *Proceedings of SemEval 2015*.

Marius Popescu. 2011. Studying translationese at the character level. In *Proceedings of RANLP 2011*, pages 634–639.

Sanja Štajner and Marcos Zampien. 2013. Stylistic changes for temporal text classification. In *Proceedings of TSD 2013*, pages 519–526.

Constantina Stamou. 2008. Stylochronometry: Stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, 23(2):181–199.