Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference

Automated Non-Content Word List Generation Using hLDA

Wayne Krug and **Marc T. Tomlinson**
Language Computer Corporation
Richardson, TX

Abstract

In this paper, we present a language-independent method for the automatic, unsupervised extraction of non-content words from a corpus of documents. This method permits the creation of word lists that may be used in place of traditional function word lists in various natural language processing tasks. As an example we generated lists of words from a corpus of English, Chinese, and Russian posts extracted from Wikipedia articles and Wikipedia Wikitalk discussion pages. We applied these lists to the task of authorship attribution on this corpus to compare the effectiveness of lists of words extracted with this method to expert-created function word lists and frequent word lists (a common alternative to function word lists). hLDA lists perform comparably to frequent word lists. The trials also show that corpus-derived lists tend to perform better than more generic lists, and both sets of generated lists significantly outperformed the expert lists. Additionally, we evaluated the performance of an English expert list on machine translations of our Chinese and Russian documents, showing that our method also outperforms this alternative.

Introduction

Non-content words are a key feature for many tasks in natural language processing (NLP). Frequently referred to as function words, these words are characterized by their lack of expression of content. They generally serve grammatical roles to link together content words, but can also be used to modify content words, such as to express doubt or negation. In languages like Chinese that lack word inflection, function words also include tense markers, collective word modifiers, and voice constructs. Function words in English include articles (e.g. the, a), pronouns, particles (e.g. if, then, however) and many other, primarily closed, word classes.

Different media have different sets of function words as well. Speech, for instance, contains various discourse markers. Some of these markers are non-lexical “filler” utterances like “uh” and “erm”. Others are words drawn in from various open word classes, such as words “well” and “like” and phrase chunks “I mean” and “on the other hand”. They serve to either connect a phrase to what comes before or after, or express an attitude of the author (Swan 2005). These discourse markers are typically edited out of prepared documents, but appear frequently in transcripts and other literal dictations of speech.

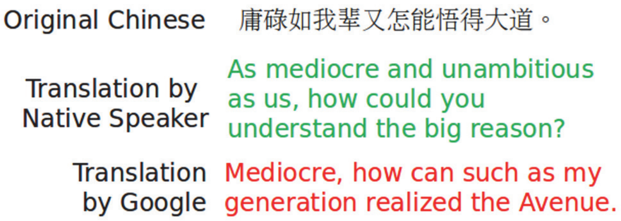
Electronic text communications have their own type of discourse markers, abbreviations such as “lol” and “imho”. Some, e.g. “imho”, are simply abbreviations of discourse markers that would otherwise be spelled out. Others, e.g. “lol”, are text representations of actions, signals, or nonverbal utterances that would normally be seen or heard in regular conversation. They are just as removed from the lexical meaning of the messages as their counterparts in speech.

In this paper we present a method for automatic discovery of non-content words from a representative corpus. This method is intended to pick up features like those described above that are typically not found on function word lists generated for conventional documents, as well as picking up the unique function word features of various languages. We demonstrate the performance in English, Russian, and Chinese of these auto-generated lists in the task of author identification against both expert-created generic function word lists and top-N frequent word lists. We also evaluate the performance of the English function word list on machine translated versions of the Russian and Chinese documents.

Background

Function words are used unconsciously and capture the stylistic choices and attitude of the author (Stamatatos 2009). They are a heavily studied feature for detecting authorship (Juola 2007), and their use dates back to the seminal Mosteller and Wallace (1964) study on *The Federalist*. They have also been shown to be useful for identifying the genre of a work as well as developing profiles of individuals (including age, gender, cognitive state, intelligence or personality) (Rude, Gortner, and Pennebaker 2004; Pennebaker, Mehl, and Niederhoffer 2003). Due to their topic independence and structural usage, function words are often amongst the most frequent in the corpus, and for this reason, many researchers use most frequent word lists as a stand-in (Juola 2007; Stamatatos 2009).

One of the issues with function words is they are dependent on language; they are sometimes also dependent on domain and medium (e.g. speech, Twitter). Normally, creating these lists requires language-specific expertise. Stamatatos (2009) mentions several efforts to define English function



Original Chinese 庸碌如我輩又怎能悟得大道。

Translation by Native Speaker As mediocre and unambitious as us, how could you understand the big reason?

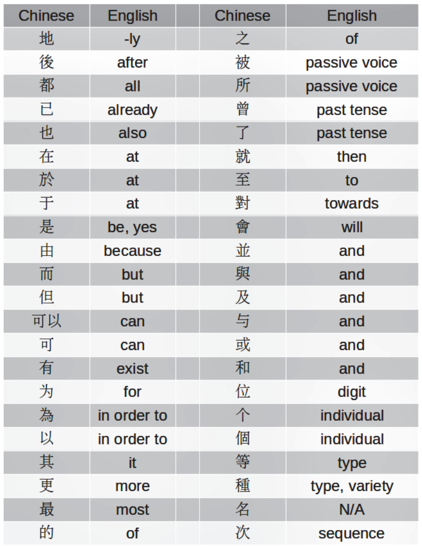
Translation by Google Mediocre, how can such as my generation realized the Avenue.

Figure 1: Example showing a Chinese sentence translated to English by both a native speaker and Google's translation service.

word lists, though he cautions that "limited information was provided about the way they have been selected". While many of these expert-created lists have been published (Stamatatos 2009), the current literature is heavily biased towards English language documents, with most work in other languages using frequent word counts instead. Some non-English function word lists have been published, in particular a list of 35 Chinese words by Yu (2012) and a list of 159 Russian words used in Snowball's Russian stemmer (Porter 2012).

As previously mentioned, many researchers work around the lack of function words in certain languages by substituting frequent word lists. Alternately, the results of Caliskan and Greenstadt (2012), during their research into the efficacy of using machine translation to obfuscate authorship features, suggest that the use of English function words, among other features, on machine translated non-English documents is effective. This would provide a second alternative. However, our research into cross-lingual authorship analysis features suggested that these alternatives would be insufficient or ineffective in language groups with significantly different syntactic and morphological structure than English.

Figure 1 illustrates some of the problems inherent in applying English features to translations. It also highlights some of the significant differences between English and Chinese language structure that complicates efforts to develop and apply cross-lingual features to these languages. Not only is the grammatical structure different, but some words simply do not translate, which leads to the poor performance of the machine translation. Some of these lost words and characters hold useful information that has no corollary in English. The chart in Figure 2 shows a number of Chinese function words and their English equivalents. Several Chinese words map to English grammatical constructs or inflections, such as the past tense characters. In addition, several English function words have multiple equivalents in Chinese. In particular, the word "and" has at least six mappings. All of this linguistic richness is lost in translation, even if the translation is performed by an expert in both languages.



Chinese	English	Chinese	English
地	-ly	之	of
後	after	被	passive voice
都	all	所	passive voice
已	already	曾	past tense
也	also	了	past tense
在	at	就	then
於	at	至	to
于	at	對	towards
是	be, yes	會	will
由	because	並	and
而	but	與	and
但	but	及	and
可以	can	与	and
可	can	或	and
有	exist	和	and
为	for	位	digit
為	in order to	个	individual
以	in order to	個	individual
其	it	等	type
更	more	種	type, variety
最	most	名	N/A
的	of	次	sequence

Figure 2: Selected list of Chinese function words with their English equivalents.

Model

The shortcomings of the function word alternatives mentioned in the previous section drove us to pursue another means of identifying function words or non-content words that operate as such. We came to the idea of the *inverted topic model*. The inverted topic model turns the typical objective of topic modeling on its head: rather than try to identify the topics and keywords with the narrowest scope, it tries to identify those with the broadest scope. These words would be non-content in the sense that they do not distinguish the content of one document from another. They are not necessarily devoid of lexical meaning. For example, if processing a corpus of scientific papers the word "science" might be determined by the inverted topic model to be non-content. "Science" clearly is a content word in the absolute sense, but becomes generic because it is applicable to the content of every document in the corpus.

This model also has value in dealing with languages like Chinese. Chinese is considered to have relatively few function words, with such words frequently omitted (Fung 1995). Chinese also does not use inflections to express different grammatical categories. Instead, it uses separate morphemes to express categories like gender, number, and tense. Researchers like Fung and Yu typically do not include these morphemes in the category of function words, since the linguistic concept does not exist in many languages. Our research indicates, however, that these language-use morphemes can be significant features for tasks such as authorship analysis; our model correctly identifies their functional use in Chinese, Japanese, and other languages.

Algorithm

We create the inverted topic model by first building a conventional topic model using hierarchical Latent Dirichlet Allocation (hLDA) (Blei et al. 2004), an extension to Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). LDA learns probability distributions over words which identify the likelihood of a word occurring in a particular topic, forming topics based on word co-occurrence within a doc-

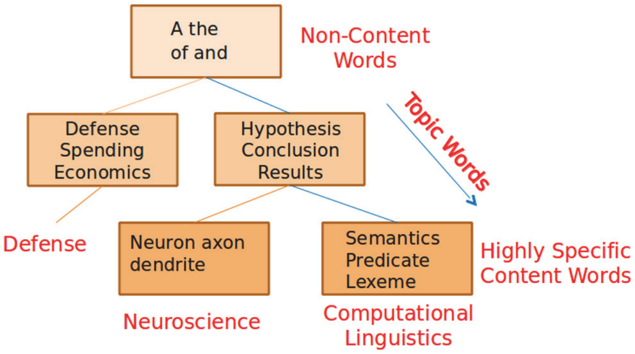

A the
of and
Non-Content Words
Topic Words
Defense Spending Economics
Hypothesis Conclusion Results
Defense
Neuron axon dendrite
Semantics Predicate Lexeme
Highly Specific Content Words
Neuroscience
Computational Linguistics

Figure 3: Example diagram showing a completed hLDA hierarchy covering a corpus of conference papers.

ument. hLDA extends this by creating a tree representing a topic hierarchy wherein each node of the tree is a topic, represented as a probability distribution across words. The root node is a topic that applies to all documents in the corpus. Subsequent branch nodes represent increasingly narrow topics, and the leaf nodes represent the most specific. Thus, an hLDA topic tree represents a partitioning of documents based on common words.

Figure 3 illustrates a possible hierarchy resulting from running the algorithm over a set of conference papers. In this case, the tree contains three leaf topics, one inner branch topic, and the root. The words in each box are the keywords extracted from the documents that describe the documents below in the tree. Thus, in the second level of the tree the word "Defense" is considered to be descriptive only of defense-related papers, while the word "Hypothesis" is considered to be descriptive of both neuroscience and computational lingustics papers.

Note that, in spite of its position, the word "Hypothesis" can appear in or describe some defense-related papers; the word's location indicates it is a much better descriptor of science papers at that particular level of the hierarchy. Similarly, some of the non-content words may reappear lower in the tree if they become descriptive of subsets of documents at that lower level. In general, the list of words associated with each topic will be ordered by the probability that those words are useful for describing all of the documents under that topic, that is, all documents that can trace a path through the tree via that node. All words with non-zero probability of describing a document may occur in a node, even if they appear in a higher node with a higher probability.

hLDA uses a nested Chinese Restaurant Process (CRP) to populate its hierarchy (Blei, Griffiths, and Jordan 2010). The nested CRP has a preset number of levels l and an infinite number of paths from level 0 (root) to level $l-1$. The algorithm runs for a set number of iterations. In each iteration, this process randomly selects one of the infinite topic paths through the levels of the topic tree for each document, and then samples the topics along that path for each word in the document. Document order is randomly shuffled periodically prior to processing the iteration. This is more formally stated in Algorithm 1. The output of the inverted topic model is the word list contained in the root node of the hLDA tree.

```
Data: A corpus of documents
A GEM distribution $GEM(m,\pi)$ (Pitman, 2002)
Result: Tree-based hierarchy of keywords organized by
        applicability of the keywords to the child
        documents
for number of iterations do
    if permutation iteration then
        Shuffle documents;
    end
    foreach document d in corpus do
        Draw path $c_d$ though a nested CRP
        Draw distribution along the levels in the tree,
        $\Theta_d|m,\pi\ GEM(m,\pi)$
        foreach word in d do
            Choose level $Z_{d,n}|\Theta_d\ Discrete(\Theta_d)$
            Draw word from each topic in path
            $W_{d,n}|Z_{d,n},c_d,\beta\ Discrete(\beta_{c_d}[Z_{d,n}])$
        end
    end
end
```

Algorithm 1: Basic algorithm for constructing the hLDA hierarchy (Blei, Griffiths, and Jordan 2010)

As evident from Algorithm 1, the runtime complexity is based on the number of iterations to run, the number of documents in the corpus, and the number of words in each document. This complexity is the algorithm's greatest weakness. Fortunately, for the purpose of non-content word generation, it need only be run once on a given corpus.

Corpus

For our evaluation we chose to use a corpus of documents extracted from the Wikipedia data dumps. Wikipedia provides a significant collection of documents in multiple languages, written by thousands of authors. These dumps consist of both the collaboratively-edited Wikipedia articles and the Wikitalk discussion pages. The Wikipedia articles provide a good representation of typical usage for their respective languages.

The Wikitalk pages consist of loosely-threaded discussions about issues with specific Wikipedia articles. In general, these postings can be tied to a specific Wikipedia contributor, making this corpus usable for authorship analysis. The loose structure provides some challenges, though, in that the onus of properly signing and delimiting posts within the thread is on the author. This leads to some parsing issues that result in either throwing out posts or combining posts from two or more authors together. In most cases, even humans cannot properly separate and identify these misformatted posts. Even so, the Wikitalk corpus represents one of the most comprehensive open-source collections of author-attrributed documents available, and utilizing a sufficiently large sample minimizes the effects of these flaws. Additionally, Wikitalk pages in all languages follow the same structure, with the same flaws, removing this as a variable in cross-lingual analysis.

For the evaluation we chose to use documents from the English, Chinese, and Russian Wikipedias, as this set of languages provides a good cross-section of language types on which to test our method. As our evaluation utilizes authorship attribution to test the effectiveness of the hLDA-generated function word lists, our test corpus is derived exclusively from the Wikitalk pages. We generated word lists from both the Wikipedia articles and Wikitalk posts.

Evaluation

The research effort that generated this algorithm was focused on identifying cross-lingual features for authorship identification, of which non-content word lists were one of the features we evaluated. Thus, our evaluation here uses the authorship identification task to assess the performance of our algorithm. While our full research used a combination of features, including hLDA and frequent word lists, for this task, the evaluation presented here uses only the word lists. This is a less-sophisticated use of function words than is typical of authorship attribution tasks; however, it is a sufficient setup to test the efficacy of our lists in comparison to the alternatives.

The objective of our evaluation was to demonstrate the performance of hLDA-generated function word lists against the alternatives, those being frequent word lists, expert word lists, and machine translations. We also wanted to test the performance of hLDA and frequent word lists as a function of list length to determine if either method would be beneficial for extending existing expert word lists. To this end we created lists of multiple lengths from our corpora.

We created four sets of word lists, one hLDA-generated function word set and one frequent word set from each of the Wikipedia article corpus and Wikitalk corpus described in the next section. Each set contains lists of 1000, 500, 200, and 100 words, as well as one matching the length of the expert list (307 for English, 159 for Russian, and 35 for Chinese). This provides multiple points of direct comparison between the function and frequent words, as well as between the languages themselves. The lists with lengths matching the expert lists allow us to test the effectiveness of a corpus-specific list without the variable of list length.

We generated function and frequent word lists using random 4000-document samplings of both the Wikitalk pages and Wikipedia articles. This provides a set of lists more reflective of the language as a whole, though still somewhat specialized (the Wikipedia articles), and a set of lists derived from documents representative of the test corpus. The minimum article or post length to be included in the list generation was 30 words. The Chinese documents were automatically word segmented, and no stemming or lemmatization was performed. Any word that occurred less than 10 times was removed from the dataset. The hLDA model was run on the data sets with the same parameters across all of the languages. The top 1000 words in the hLDA root node and input histogram became the base lists for the function word and frequent word features respectively.

To perform the authorship attribution task we created a model in the Weka machine learning system (Hall et al. 2009) using the J48 decision tree algorithm. The lists of words formed the feature vectors for each trial. The feature magnitudes were the normalized frequency of the corresponding word in the document, i.e. word density. To evaluate the quality of the word lists and densities as feature vectors in the model we ran Weka's 10-fold cross validation. This was run as a multi-class classification problem, that is, the model as trained with samples from all 100 authors in each corpus, and had to choose between all 100 when making a judgment as to the authorship of a particular document. All performance results presented later are the accuracies reported by the Weka cross-validation runs.

Language	Authors	Documents
English	100	2781
Russian	100	3997
Chinese	100	3549

Table 1: Number of posts yielded from the 4000 selected for each language

The evaluation data was created by parsing posts from the Wikitalk pages in the data dumps for English, Russian, and Chinese. This corpus contains several thousand identifiable authors, whose posts range in length from a few to several thousand words. We randomly selected 40 posts each from the top 100 authors without regard to post content, for a nominal total of 4000 posts per language. These posts were stripped of e-mail addresses and URLs, then turned into documents. However, due to the quality of the parse and the nature of some of the posts, we could not successfully convert them all into documents. Table 1 shows the yield for each language from the 4000 selected posts. The generated test corpus is representative of a real-world dataset, providing a good, practical test for our method. We were concerned about biasing the results if we spent too much effort cleaning and massaging the dataset.

Results

Our first test was to compare the expert lists we have for the three languages (Yu 2012; Porter 2012; Zeman 2011) to an hLDA list and frequent word list of the same size. Table 2 shows the results. Both the hLDA and frequent word lists significantly outperform the expert lists in all languages, for both the Wikipedia and Wikitalk variants. In all cases, however, the frequent word lists hold a slight advantage over the hLDA lists.

To further explore the function vs. frequent word question we compared the five sizes of hLDA function word lists to corresponding size frequent word lists. Table 2 shows the results of those trials. A few conclusions are evident from this table. First, the frequent word list consistently holds a slight advantage over the hLDA lists for all language and list lengths, with the most significant being in Chinese. Second, there is a distinct advantage to increasing the size of the list, regardless of the source data. Third, corpus specific lists have a slight advantage over more generic language lists, with the exception of the shorter Russian lists.

Language	List Length	Expert	Wikipedia		Wikitalk	
			hLDA	Frequent	hLDA	Frequent
English	100	N/A	11.506	11.754	9.660	14.773
	200	N/A	12.429	13.210	14.773	16.690
	307	5.575	14.276	17.294	16.903	18.892
	500	N/A	16.974	18.679	19.211	25.249
	1000	N/A	20.171	20.668	30.753	31.285
Chinese	35	4.170	10.341	13.328	6.650	12.933
	100	N/A	13.102	13.948	9.467	15.976
	200	N/A	13.750	15.835	10.285	18.738
	500	N/A	15.751	17.949	13.948	19.301
	1000	N/A	16.765	19.104	20.006	20.795
Russian	100	N/A	5.679	6.530	5.204	5.804
	159	1.826	6.205	6.680	5.054	6.155
	200	N/A	6.830	6.780	5.579	7.205
	500	N/A	6.355	6.955	7.055	7.531
	1000	N/A	6.355	8.081	8.056	9.882

Table 2: Comparison of hLDA function words, frequent words, and expert function words for three languages.

The performance of the hLDA lists was lower than expected, with the frequent word lists perfoming better by anywhere from 0.751 points to 5.478 points on average across the list sizes. Because of this, we compared the overlap in the words between each pair of lists. As mentioned previously, function words are expected to be among the most frequent words in a particular set of documents. Table 3 shows the average overlap between the hLDA function word and frequent word lists for each language and source corpus. Most interesting is that, while the lists derived from the Wikitalk corpus have little overlap, their performance on the corpus is still comparable. This suggests the hLDA model identified some words useful to the authorship attribution task that are relatively infrequent in the corpus. However, there does not appear to be a correlation between the similarity of the lists and the delta in performance between them.

Since the lists created off the Wikitalk post corpus, which is representative of the data on which we tested, performed noticeably better in most cases than the lists generated off the Wikipedia article corpus, we wanted to see how similar the word lists were between the corpora. Thus, we also calculated the average overlap between each type of list for each source corpus, with the results given in Table 4. It is clear from this table that the two corpora differ significantly, particularly in what the hLDA model considers to be non-content words, further reinforcing the postion of using representative samples when possible. Once again, though, there does not appear to be a correlation between the similarity ratios and performance deltas.

Language	Wikipedia	Wikitalk
English	0.623	0.241
Chinese	0.470	0.362
Russian	0.682	0.291

Table 3: Average overlap ratio between frequent and function word lists for each language and list source corpus.

Language	Function	Frequent
English	0.173	0.420
Chinese	0.229	0.338
Russian	0.194	0.485

Table 4: Average overlap ratio between the lists generated from the wikitalk corpus compared to those generated from the wikipedia corpus for lists based on frequent words or function words.

As previously noted, our tests use the word lists and associated word densities as feature vectors in a machine learning classifier, applied to the task of authorship attribution. This is a simplistic approach to the task, which is the reason the absolute numbers are so low compared to most published results. They key takeaway from the results presented in this section is not the absolute performance of the lists but rather the relative difference, or lack thereof.

Finally, we analyzed the effectiveness of using English function words to identify authors in machine translated documents. We utilized Microsofts Translate service[1] to translate our Chinese and Russian datasets into English. We then used the English expert list on the translated documents in the same manner as on the native English documents. The results of these trials are in Table 5. "Native" accuracy for Chinese and Russian on this table is for the appropriate 200-word hLDA list. Here the hLDA lists outperform the alternative of machine translating the documents and using the English list. While the Russian performance is on par with its expert list, the 35-word Chinese expert list once again outperforms the alternative.

Conclusion and Future Work

Overall hLDA lists performed on par, or slightly worse, than the lists based on frequent words at identifying the authorship of wikitalk pages. Both approaches significantly out-

[1] http://www.microsofttranslator.com

Language	Native	Translated
English	5.575	N/A
Chinese	13.750	2.313
Russian	6.830	2.052

Table 5: Comparison of native language function word lists to English expert list in machine translation.

performed the lists generated by experts. The deficit in performance for the expert created lists is surprising given the continued reliance on function word lists in some communities.

The lists generated off of the corpus representative of the test documents performed slightly better in almost all cases than the lists generated off a more generic corpus. This result reinforces the value of tailoring word lists to the target corpus, which is one of the objectives of our method.

For the translations, accuracy of the translation will greatly affect the distribution of the words in the resulting documents. Furthermore, different languages have different sets of non-content words, so even a perfectly accurate translation would not properly capture the author's profile. However, this loss of information could be combined with the work done by Caliskan and Greenstadt (2012) to assist in identifying translated works and the source of the translation.

Overall the inverted topic model based on hLDA showed a good ability to pick out non-content words that were not just the most frequent words. While its native performance for the particular task of authorship identification does not show an increased effectiveness over frequent word based approaches the model holds considerable promise for other tasks. Additionally, because the function and frequent word lists had so little overlap, but still performed similarly indicates that the hLDA list and the frequent word list contain information which would be mutually beneficial and could be exploited in further work.

We would also like to further study the hLDA-based inverted topic model to determine if any other useful information can be extracted from it. The model is far more than just a collection of topics and words; it contains cooccurence statistics, word distributions, and other information that could be exploited in ways beyond just developing lists of discriminatory and non-discriminatory words.

Further, in the introduction we mentioned the peculiar discourse markers of speech and electronic text communication. However, we focused this study on the language differences. We would like to repeat our analysis on documents drawn from these two media to evaluate the inverted topic model's effectiveness at identifying their discourse markers, and the effectiveness of using them as function words.

Acknowledgements

This material is based upon work supported by the Office of Naval Research under Contract No.N00014-12-M-0286.

References

Blei, D.; Griffiths, T.; Jordan, M.; and Tenenbaum, J. 2004. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*, volume 16, 17. MIT Press.

Blei, D.; Griffiths, T.; and Jordan, M. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)* 57(2):7.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Caliskan, A., and Greenstadt, R. 2012. Translate once, translate twice, translate thrice and attribute: Identifying authors and machine translation tools in translated text. In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on*, 121–125. IEEE.

Fung, P. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Proceedings of the 3rd Workshop on Very Large Corpora*, volume 78.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter* 11(1):10–18.

Juola, P. 2007. Authorship attribution. *Foundations and Trends® in Information Retrieval* 1(3):233–334.

Mosteller, F., and Wallace, D. L. 1964. Inference and disputed authorship: The federalist.

Pennebaker, J.; Mehl, M.; and Niederhoffer, K. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54(1):547–577.

Pitman, J., et al. 2002. Combinatorial stochastic processes. Technical report, Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for St. Flour course.

Porter, M. 2012. Snowball: A language for stemming algorithms. http://snowball.tartarus.org/.

Rude, S.; Gortner, E.; and Pennebaker, J. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* 18(8):1121–1133.

Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60(3):538–556.

Swan, M. 2005. *Practical English Usage*. Oxford University Press.

Yu, B. 2012. Function words for chinese authorship attribution. In *Computational Linguistics for Literature, Workshop on*, 45–53. ACL.

Zeman, D. 2011. Message to universitetet i bergen corpora mailing list. http://comments.gmane.org/-gmane.science.linguistics.corpora/14448.