# Deep Level Lexical Features for Cross-lingual Authorship Attribution.

Marisa Llorens-Salvador. Sarah Jane Delany.

Dublin Institute of Technology, Dublin, Ireland

**Abstract.** Crosslingual document classification aims to classify documents written in different languages that share a common genre, topic or author. Knowledge-based methods and others based on machine translation deliver state-of-the-art classification accuracy, however because of their reliance on external resources, poorly resourced languages present a challenge for these type of methods. In this paper, we propose a novel set of language independent features that capture language use from a document at a deep level, using features that are intrinsic to the document. These features are based on vocabulary richness measurements and are text length independent and self-contained, meaning that no external resources such as lexicons or machine translation software are needed. Preliminary evaluation results show promising results for the task of crosslingual authorship attribution, outperforming similar methods.

**Keywords:** Crosslingual document classification, crosslingual authorship attribution, deep level lexical features,vocabulary richness features.

## 1 Introduction

Despite the prevalence of the English language in many fields, international organizations manage large numbers of documents in different languages, from local legislation to internal documents produced in different company locations. At the same time, workers' mobility has created a multilingual work force that create and store documents in different languages depending on the context. For example, the same author can write academic papers in English, write a technical book in French and a novel in Catalan. The classification of these multilingual documents has applications in the areas of information retrieval, forensic linguistics and humanities scholarship.

The analysis of document style and language use has long been used as a tool for author attribution. Traditionally, research in the area focused on monolingual corpora [12] or employed external resources such as machine translation, multilingual lexicons or parallel corpora [3, 14, 15].

In this paper, we present a set of language independent lexical features and study their performance when used to solve the problem of crosslingual author attribution. The task of crosslingual author attribution (CLAA) refers to the

identification of the author of a document written in language $x_i$ from a pool of known authors whose known documents are written in languages $x_1, x_2, .., x_n$. The aim of the method is to identify the author of an unseen document without prior knowledge about its language, i.e. without using any language specific features, tuning for a particular language or the use of machine translation/lexicon aid in a completely language independent implementation.

The proposed method builds on traditional vocabulary richness measures (VR), such as type-token ratio or hapaxes frequency. Traditional vocabulary richness features are text-length dependent and provide a small number of features (type-token ratio being the best example with only one value representing each text). In order to overcome these limitations, our proposed method for feature extraction calculates features on fixed length samples of text extracted from the document. Mean and dispersion values for vocabulary richness are calculated obtaining 8 deep level lexical features. The performance of different sample sizes $i$ is studied individually and as combinations of sizes, providing information about text consistency through the document and characteristic vocabulary use.

## 2 Related Work

Monolingual author attribution has in the last few years achieved a high level of accuracy using lexical features such as frequencies of the most common words and Burrow's Delta to calculate distances between documents [1, 4, 7, 11, 13]. Other lexical features used in monolingual author attribution include frequencies of stop words [2] and word n-grams. In these models, a feature vector with all features (n-grams or stop words) contained in the document and their frequencies characterizes each document. The problem when extending these methods to multilingual corpora is that the dimensions of the feature vectors in different languages are in general orthogonal, giving zero as the similarity measure between documents. Character n-grams have been applied to different languages and have obtained high levels of accuracy at the expense of high dimensionality with feature set sizes in the thousands [7]. At a syntactic level, features such as part-of-speech and frequency of verbs and pronouns have achieved high level of accuracy as well [6]. However, all the above features are either language dependent or involve high dimensional feature sets.

Traditional vocabulary richness like the type-token ratio are language independent, however, they depend on text length and for this reason have been replaced in recent times by more complex features. These features include the Moving Window Type-Token Ratio and the Moving Window Type-Token Ratio Distribution [5, 8]. Despite their language independence nature, traditional measurements of vocabulary richness have not delivered highly accurate results in the past [13]. Consequently, they have been replaced by the use of lexical features in combination with machine translation software or lexicons/dictionaries

to bring all documents into the same language space with *wikipedia* and the *eurovoc corpus* the most commonly used resources [9, 10, 14].

## 3 Methodology

Based on vocabulary richness and frequency spectrum values, the proposed features and method for feature extraction define a way of quantifying the style of a text by analysing the use of vocabulary in samples of different sizes taken from the text. These samples are based on the idea of a moving window type-token ratio using fixed size samples and hence avoiding the shortcomings of the type-token ratio. These features extend the moving window type-token ratio as more granular measurements of word frequencies are extracted.

Three sampling methods are included in the framework: (i) Fragment sampling (FS), (ii) Bags-of-words sampling (BS) and (iii) the combination of both Bag-Fragment sampling (BFS).

Fragment sampling (FS) is defined as the process of randomly obtaining $n$ samples of $i$ consecutive words, starting from a word chosen at random and each sample is referred to as a fragment. Given the random nature of the sampling process these fragments can overlap and are not following any sequence in terms of overall location in the text. Bags-of-words sampling (BS) involves the use of $i$ words sampled randomly from any part of the document and follows the well known concept of treating a text as a bag-of-words where the location of words is ignored .

The proposed set of language independent lexical features is extracted following a 4 step process:

STEP 1: A number $n$ of document samples of size $i$ is extracted.
STEP 2: Values for frequency features are calculated per sample.
STEP 3: Values for mean and dispersion features calculated across the $n$ samples.
STEP 4: Back to step 1 for a new sample size $i$.

The general parameters of the method are: type of sample (Fragment, Bags-of-words or both), sample sizes $i_1$,$i_2$,...,$i_M$ and number of samples $n$ per sample size. Figure 1 depicts a diagram for the extraction process for BFS. FS and BS are represented by the left and right hand-side of the diagram respectively.

The proposed set of frequency features are based on the analysis of the frequency spectrum, i.e. how many times each feature appears. A typical example of this type of features is the number of hapaxes or words that appear only once in the text. Instead of using the entire frequency spectrum and in order to reduce the number of features and capture information in a compact way, a novel

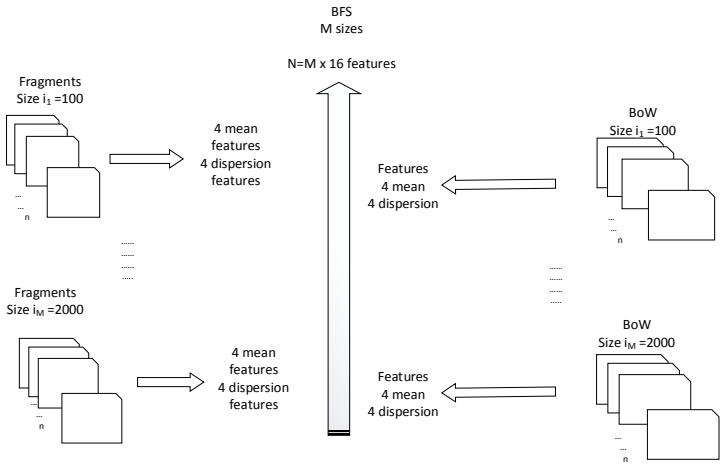


**Fig. 1.** BFS process summary diagram.

method of frequency spectrum representation is presented.

The frequency spectrum for different texts shows regular behaviour for the initial low frequencies, however, after frequency 10 the number of words for each frequency becomes less stable as can be seen in Figure 2, which shows the frequency spectrum for Charles Dickens' Oliver Twist in its original language. For this reason, frequency values over 10 are not used for the purpose of feature extraction. Notwithstanding these considerations, the words included in that frequency range (over 10) are not entirely neglected as they feature as part of the overall vocabulary and hence contribute to the classification process.

The frequency spectrum for values of frequency between 1 and 10 is regular (quasi linear) and hence suitable for a small number of points to represent its behaviour. In order to reduce the dimensions of the feature set and given the quasi linear behaviour of the data, a further simplification is performed and groupings of 1, 2-4, and 5-10 are used. Each frequency range is represented by a feature, obtaining 3 features to represent the frequency spectrum between 1 and 10 and a separate fourth feature that represents the vocabulary or different unique words present in the text. Figure 2a shows the 3 features representation of data for Charles Dickens' Oliver Twist in its original language English plotted on top of the overall frequency spectrum.

The feature representation of the frequency spectrum for values of frequency between 1 and 10 holds for fragments and bags-of-words samples as shown on Figure 2b. The sampling process allows for dispersion features to be calculated

providing a measurement of the homogeneity of the text.

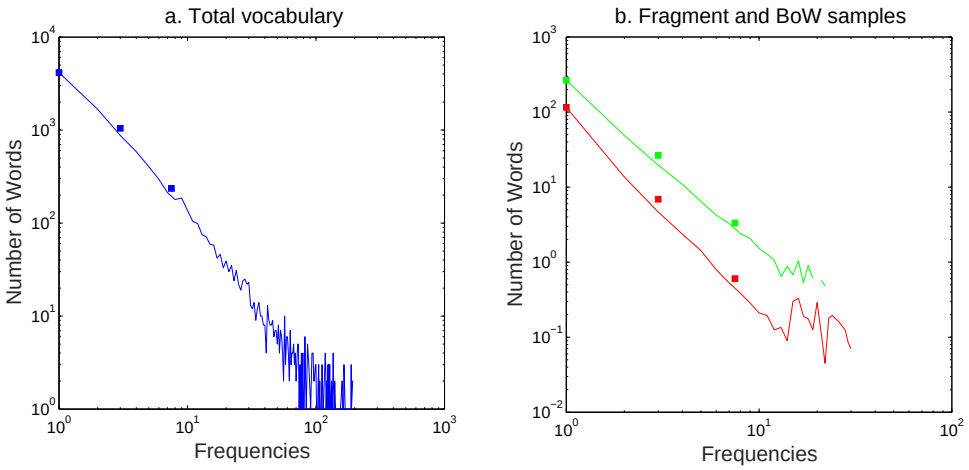


**Fig. 2.** Oliver Twist (Charles Dickens) a. Frequency spectrum with 3 features b. Fragment and bags-of-words sample (i=200) with 3 features.

Table 1 shows the proposed mean and dispersion features for the frequency groupings and vocabulary.

| | |
|---|---|
| 1 | Size of vocabulary per sample. |
| 2 | Number of local hapaxes $h_i$. |
| 3 | Number of words with frequency 2, 3 and 4. |
| 4 | Number of words with frequency 5 to 10. |
| 5. | Coefficient of variation for vocabulary. |
| 6. | Coefficient of variation for local hapaxes. |
| 7. | Coefficient of variation for words with frequency 2, 3 and 4. |
| 8. | Coefficient of variation for words with frequency 5 to 10. |

**Table 1.** Deep level features.

The sampling process is repeated for a number, $M$, of sample sizes, $i$, and the 8 features calculated for each size. This provides a variable number of final features depending on the number of sizes selected. The size of the resulting set of features depends on $M$, the number of different sizes sampled. The total number of features $N$ is $N = 8 \times M$ for FS and BS and $N = 16 \times M$ for BFS.

### 3.1 Datasets

In order to adjust the parameters of the proposed feature extraction method, a multilingual corpus of literary works was compiled. Due to the cross-lingual nature of the experiments, documents in different languages created by the same author are required. Literary translation is believed to keep the markers from the original author and the influence of the translator is weak [16], therefore the corpus used in the experiments is formed by original works by 8 authors and translated novels from the same 8 authors. It includes two datasets: Dataset 1, a balanced dataset of original documents and Dataset 2 a unbalanced extended version including translations. Dataset 1 contains 120 literary texts from 8 different authors (15 documents per author) in 4 different languages (English, Spanish, German and French) as can be seen in Table 2. Dataset 2 includes all documents from Dataset 1 plus 85 additional documents which are translations of literary texts by some of the 8 authors from Dataset 1. A summary of the translations in Dataset 2 can be found in Table 3. All documents were obtained from the Gutenberg project website[1].

| **Language** | **Author** | **Average document length** |
|---|---|---|
| English | Charles Dickens | 144222 |
| English | Ryder Haggard | 97913 |
| French | Alexander Dumas | 139681 |
| French | Jules Verne | 84124 |
| German | J. W. von Goethe | 67671 |
| German | F. Gerstäcker | 51655 |
| Spanish | V. Blasco Ibañez | 100537 |
| Spanish | B. Perez Galdos | 126034 |

**Table 2.** Dataset 1 description: 4 languages, 8 authors and 15 documents per author.

| **Author** | **Language (# documents)** |
|---|---|
| Charles Dickens | French (13) |
| Alexander Dumas | English(19) Spanish (2) |
| Jules Verne | English (21) German (3) Spanish (1) |
| J. W. von Goethe | French (1) English (6) |
| V. Blasco Ibañez | English (13) French (2) |
| B. Perez Galdos | English (5) |

**Table 3.** Dataset 2 description: language and number of translated documents.

[1] https://www.gutenberg.org/

### 3.2 Estimating optimum parameter values

The first parameter to be set is $n$ the number of samples for each sample size $i$ that is necessary to obtain a representative figure for average and dispersion values. An empirical study has been performed with 10 to 2000 samples of each size, using a Random Forest classifier and *leave one out* cross validation. The results of the classification using Fragments and bags-of-words for Dataset 1 are shown on Figure 3.

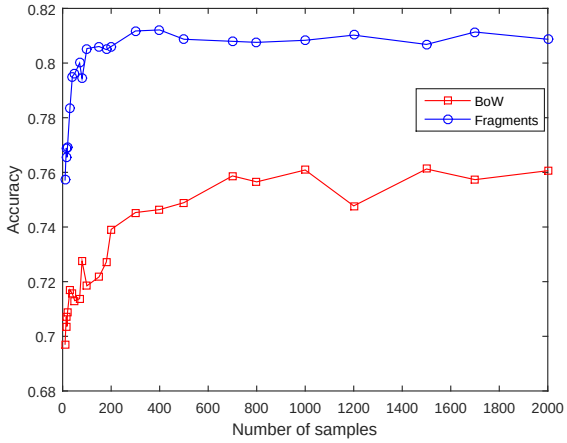


**Fig. 3.** Number of samples vs. correctly classified documents

The number of correctly classified documents increases as the number of samples increases until a stable value is reached. Fragments and bags-of-words behave differently with more variation in the bags-of-words samples. Two threshold levels can be identified in figure 3, the first threshold is around the value of 200 samples, and the second threshold is around 700 samples where the results are more stable. However, as the computational time is an important factor in text analysis, the selected value for n, the number of samples, is fixed at 200 samples per sample size $i$.

### 3.3 Optimum sample size or combination of sample sizes. Number of features.

Once the number of samples is fixed, we need to determine the sample sizes $i$ that will produce the best performing set of features. For each sample size, the proposed method produces a set of 8 features. All sample sizes and their combinations will be empirically tested to evaluate the effect of different numbers of features on the final classification. For this experiment, the following sample

sizes (fragments and bags-of-words) have been used: 200, 500, 800, 1000, 1500, 2000, 3000 and 4000.

Combinations of 1, 2, 3, 4, 5, 6, 7 and 8 different sample sizes were taken for both fragment and bag-of-words samples, as well as the combination of both types of samples. In order to optimize the number of features, the combination that produces the highest accuracy with the lowest number of features will be selected.The results, grouped per number of different sample sizes ($M$) and hence per total number of features, are shown in Figure 4. Figure 4 shows the results for fragments, bags-of-words and the combination of both for Datasets 1 and 2.

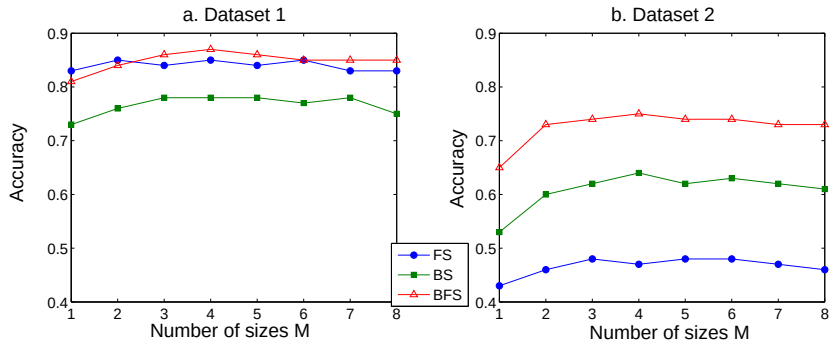


**Fig. 4.** Accuracy FS, BS and BFS for Datasets 1 and 2

The results from the different combinations of sample sizes show different responses to Dataset 1 and Dataset 2. The different nature of these two datasets explain the different behaviour of the type of samples for each dataset. Fragments are more powerful at discriminating between originals in a balanced setting whereas bags-of-words perform poorly when each author is represented by documents in only one language. On the other hand, bags-of-words provide stronger results for the more difficult problem presented in Dataset 2 where translations are included in the dataset. In both scenarios, the combination of both types of samples, BFS, provides the best results.

In terms of the final size of the feature set, which combines the type of sample and the number of sample sizes $i$, there is no significant improvement after 2 sizes are combined. The final size of the feature set is therefore $N = 2(8_F + 8_B) = 32$. A closer look at the combination of sizes that produce the best results show sizes 500 and 1000 obtaining the highest accuracy.

Preliminary evaluation of BFS applied to CLAA using the same cross-validation method (*leave one novel out*) and the same dataset as Bogdanova and Lazari-

dou [3] shows that BFS achieves better classification results (0.47) than high level features without the use of machine translation (0.31). In this particular experiment, 27 documents plus 7 which are translations of one of the 27 are used, with the final dataset being formed by 275 texts extracted from the 34 original documents. For this reason, *leave one novel out* is used to avoid the classifier being trained on texts from the same document (or translations of it). Every time *leave one novel out* is performed on this dataset, a large number of texts are removed from the training data, hence the training set is small, which added to the short length of the texts, affects the overall classification performance. Machine translation methods achieve better results but are limited by the availability of resources in the given languages as well as the requirement to identify the target language beforehand.

## 4 Conclusion

This paper has presented a feature extraction method for language independent vocabulary richness measurements. Traditional vocabulary richness methods have not performed to state of the art accuracy values in the past and have been replaced with monolingual features such as word n-grams and part-of-speech features. In order to work with multilingual corpora, previous research has used machine translation [3] and lexicons or texts available in several languages such as *wikipedia* [9] or *eurovoc* documents [14]. The proposed method expands traditional vocabulary richness using two types of samples: fragments and bags-of-words of fixed size. It calculates local measurements on those samples as well as the dispersion of those measurements over the samples. The method uses solely deep level intrinsic document measurements and hence no external resources are used.

Our experiments on cross-lingual authorship attribution show that BFS with deep lexical features is suitable for discriminating between authors in multilingual task using a relatively small feature set and no external resources. Even though the accuracy of machine translation based methods is still significantly higher, the experiments reproduced deal with highly popular languages such as English and Spanish, and results for low resource languages are expected to be lower. In these situations, a method based on intrinsic document features such as the one presented in this paper, provides a solution that is not biased by the amount of external resources available. Further work will focus firstly on extensive evaluation of the performance of BFS at a variety of cross-lingual tasks and secondly on the exploration of deep level features used in combination with other language independent methods (implementation-wise) such as character n-grams or methods based on punctuation and sentence length measurements.

## References

1. Ahmed Shamsul Arefin, Renato Vimieiro, Carlos Riveros, Hugh Craig, and Pablo Moscato. An Information Theoretic Clustering Approach for Unveiling Authorship

Affinities in Shakespearean Era Plays and Poems. *PLoS ONE*, 9(10):e111445, October 2014.

2. R. Arun, Suresh V. Murty Saradha, R., and C. E. Veni Madhavan. Stopwords and stylometry: a latent Dirichlet allocation approach. In *NIPS workshop on Applications for Topic Models: Text and Beyond*, pages 1–4, 2009.
3. Dasha Bogdanova and Angeliki Lazaridou. Cross-Language Authorship Attribution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, number May, pages 83–86, 2014.
4. John Burrows, David Hoover, David Holmes, Joe Rudman, and Fiona J Tweedie. The State of Non- Traditional Authorship Attribution Studies 2010 : Some Problems and Solutions. *Source*, pages 1–3, 2010.
5. M. A. Covington and J. D. McFall. Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100, 2010.
6. Michael Gamon and Agnes Grey. Linguistic correlates of style : authorship classification with deep linguistic analysis features. *Proceedings of the 20th International Conference on Computational Linguistics*, 4:611, 2004.
7. V. Keselj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. *Computational Linguistics*, 3:255–264, 2003.
8. Miroslav Kubát and Jiří Milička. Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics*, pages 339–349, 2013.
9. Mari-Sanna Paukkeri, Ilari T. Nieminen, P. Matti, Matti Pöllä, and Timo Honkela. A Language-Independent Approach to Keyphrase Extraction and Evaluation. In *COLING (Posters)*, number August, pages 83–86, 2008.
10. Salvatore Romeo, Dino Ienco, and Andrea Tagarelli. Knowledge-Based Representation for Transductive Multilingual Document Classification. *ECIR 2015*, a:92–103, 2015.
11. Jan Rybicki and Maciej Eder. Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, 26(3):315–321, September 2011.
12. Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, March 2009.
13. Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics*, 26:471–495, 2000.
14. Ralf Steinberger, Bruno Pouliquen, and Johan Hagman. Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC. In *Computational Linguistics and Intelligent Text Processing, Third International Conference*, pages 415–424, 2002.
15. Lauren M. Stuart, Saltanat Tazhibayeva, Amy R. Wagoner, and Julia M. Taylor. Style features for authors in two languages. *Proceedings - 2013 IEEE/WIC/ACM International Conference on Web Intelligence*, 1:459–464, 2013.
16. Lawrence Venuti. *The translator's invisibility: A history of translation*. Routledge, 2008.