Cross-Language Authorship Attribution

Dasha Bogdanova[(1)], Angeliki Lazaridou[(2)]

(1) CNGL Centre for Global Intelligent Content, School of Computing, Dublin City University, Dublin, Ireland
(2) Center for Mind/Brain Sciences, University of Trento, Italy
dbogdanova@computing.dcu.ie, angeliki.lazaridou@unitn.it

Abstract

This paper presents a novel task of cross-language authorship attribution (CLAA), an extension of authorship attribution task to multilingual settings: given data labelled with authors in language X, the objective is to determine the author of a document written in language Y, where $X \neq Y$. We propose a number of cross-language stylometric features for the task of CLAA, such as those based on sentiment and emotional markers. We also explore an approach based on machine translation (MT) with both lexical and cross-language features. We experimentally show that MT could be used as a starting point to CLAA, since it allows good attribution accuracy to be achieved. The cross-language features provide acceptable accuracy while using jointly with MT, though do not outperform lexical features.

Keywords: Cross-Language Techniques, Authorship Attribution, Text Classification

1. Introduction

Authorship Attribution (AA), the task of identifying the author of an anonymous text, has a long history and various methods dealing with this task have been proposed (Stamatatos, 2009). However, none of the proposed methods found in the literature considers the scenario where the same person writes documents in different languages. Nowadays, with the fast growth of the web, users tend to participate in various online communities irrespective of the language. Focusing on social media, a researcher from Spain may have a blog in Spanish, a twitter in English and publish research papers in both languages. At the same time, various novelists write in more than one language. As an example, the Russian-American novelist Vladimir Nabokov wrote in both English and Russian, and the Irish writer Samuel Beckett wrote in English and French. Thus, we foresee a substantial need in reliable methods for cross-language authorship analysis.

Even though recently attention has been given to cross-language technologies ranging from information retrieval to text classification, to the best of our knowledge the problem of cross-language authorship attribution (CLAA) has not been addressed so far. The most related task is that of Cross-language Plagiarism Detection; methods (Potthast et al., 2011; Franco-Salvador et al., 2013) dealing with that task focus on using lexical features for languages with similar alphabets as well as features extracted from the multilingual semantic network BabelNet (Navigli and Ponzetto, 2010).

In this paper, we present the first attempt towards CLAA. Similarly to the cross-lingual literature (Wan, 2011; Shi et al., 2010), our starting point is to build a monolingual AA system in the language for which we have reliable resources and then use Machine Translation (MT) to translate any other testing data to that language. However, this is not the optimal solution. First of all, machine translation is arguably a very challenging task and even though considerable research has resulted in more accurate techniques, the state-of-the-art is still far from perfect, thus introducing a natural error propagation when used as a form of preprocessing. For this reason, as an alternative to MT, we explore stylometric features that are either cross-language or are language-dependent but can "survive" the MT. Our main contributions are three-fold: (1) we propose a novel task of CLAA; (2) we present a method to perform CLAA; (3) some of the suggested cross-language features, i.e., perceptual ones, had not been explored as features for authorship analysis.

The rest of this paper is structured as follows: Section 2 presents the task of cross-language authorship attribution, and the main approaches we consider: the approach based on cross-language features is described in Section 1.1 and the machine translation based approach is described in Section 2.1. In Section 4 we present the dataset used in the experiments. The experimental setup is described in Section 3, followed by the experimental evaluation and results in Section 5. We share the plans for future work in Section 6 and draw conclusions in Section 7.

1.1. Cross-Language Features

In monolingual AA, good performance has been achieved by using simple lexical features (Keselj et al., 2003; Peng et al., 2003) and in some cases more elaborated, though still language specific, features, such as n-grams of POS (Gamon, 2004) and syntactic tags (Hirst and Feiguina, 2007). When we face the task of CLAA, these features are no more applicable. Thus, there is a need in finding features that are both language independent and reflect the author's writing style.

In this work we employ the following categories of features: sentiment features (frequency of positive/negative words), emotional features (frequencies of basic emotions: joy, anger, fear, sadness, surprise, disgust), POS tags (frequencies of nouns, verbs, adjectives and adverbs), perceptual features (frequencies of markers of visual, audial and kinesthetic perception) and average sentence length. The features we use in this study are presented in Table 2.

Feature Category	Number of features	English Resources	Spanish Resources
Sentiment	2	SentiWordNet (Esuli and Sebastiani, 2006)	Spanish Sentiment Lexicon (Perez-Rosas et al., 2012)
POS frequencies	4	Stanford POS tagger (Toutanova et al., 2003) and opennlp POS tagger[1] then mapped to universal POS tags (Petrov et al., 2011)	
Emotions	6	WordNet-Affect (Strapparava and Valitutti, 2004)	Spanish Emotion Lexicon[2] (Sidorov et al., 2012)
Perception	3	Our approach described in Section 1.1 and MCR (Atserias et al., 2004)	
Avg. sentence length	1	Sentence splitting done with Stanford parser (Klein and Manning, 2003)	

Table 1: Features used in experiments.

Our motivation for introducing features based on sentiment and emotional features is based on the fact that they are cross-language, i.e., people experience and express emotions and sentiments irrespective of their native language. Furthermore, previous research (Panicheva et al., 2010) reports that sentiments are expressed differently by different people. We believe, expression of certain emotions as well as sentiments could be a clue to an author's style and personality. Although POS tags frequencies have been proven (Gamon, 2004) to be a helpful feature in monolingual AA, they are clearly not a cross-language feature. However, we believe that frequencies of certain POS tags, e.g. nouns, verbs, adjectives and adverbs, should remain stable after translation. The only feature category that to the best of our knowledge has never been employed before is perception category. In the next paragraph we describe how these features are extracted.

According to psychological studies (Dunn et al., 1989), most of the people could be roughly divided into visual, audial and kinesthetic learners, i.e. those who better perceive visual, audial or kinesthetic information respectively. We believe that learning style is reflected in the language one speaks. In order to extract perceptual features from documents, we compiled a list of markers for the three learning styles. We started by creating lists of seed words that indicate each learning style, e.g. *hear, listen, sound, noise, music, loud, quiet* indicate audial perception; visual: *touch, feel, smell, pain, feeling, smooth, rough, stinky, smelly*; kinesthetic: *see, watch, look, color, bright, dark, light, big, large, small, little*). Then, we used the Multilingual Central Repository (MCR) (Atserias et al., 2004) based on WordNet 3.0, to map the synsets of the English seed words to the corresponding Spanish words, and thus, constructed seed words lists for Spanish. Following that, we extended the initial set of seed words with their synonyms (i.e. words belonging in the same synset) as well as their hyponyms as extracted from the MCR. Finally, this process resulted in 58 and 54 markers of audial perception, 110 and 140 markers of visual perception, 84 and 87 markers of kinesthetic perception for English and Spanish respectively.

Since some of the features, i.e., ones based on POS frequencies, are not cross-language, we further refer to the features listed in Table 2 as the high-level features.

[1] http://opennlp.apache.org/

[2] http://www.cic.ipn.mx/~sidorov/\#SEL

2. Cross-Language Authorship Attribution

One of the main authorship analysis problems is authorship attribution (AA). Monolingual AA is usually formulated as a classification task, i.e., given training data labelled with authors, the goal is to determine the author of unseen texts. In the cross-language setting, the task is formulated in a similar way: given data labelled with authors in language X, the objective is to determine the author of a document written in language Y, where $X \neq Y$.

2.1. Machine Translation

Even though the quality of MT is not always optimal, especially for low-resource languages, we suggest to use MT as a starting point for CLAA. MT brings documents written in different languages into one space, and thus, enables the use of lexical features. However, poor MT could happen to bring its own "style" that could mask the style of the author. In this paper, apart from the cross-language features approach, we also explore MT as a way to perform CLAA.

3. Experimental Settings

Here we have formulated the task of CLAA as a classification task; given data in English labelled with authors, classify a new Spanish document according to its author. Apart from the choice of the classification method, in the task of CLAA it is important to choose a cross-language representation of the documents, i.e. the way to bring the documents to the same feature space, since they are in different languages initially. We experimentally explore the following approaches:

1. **MT + lexical features.** We translate Spanish documents to English[3], and represent documents with word, character or POS n-grams.

2. **High-level features.** We construct vector representations for the two languages separately using the features described in Section 1.1.

3. **MT + High-level features.** We first translate Spanish texts to English as in 1., and then construct vector representations with the features described Section 1.1, having all data in English.

[3] we use Google Translate: http://translate.google.com

Author	English	Spanish
Charlotte Brontë	*Jane Eyre, The Professor, Villette*	*Jane Eyre*
Rudyard Kipling	*From Sea to Sea, The Jungle Book, Captains Courageous, Kim*	*The Jungle Book, The Phantom 'Rickshaw*
Lewis Carroll	*Alice in Wonderland, Sylvie and Bruno, The hunting of the snark*	*Alice in Wonderland, Through the looking-glass*
Robert Stevenson	*Treasure Island, The black arrow, Strange case of Dr Jekyll and Mr Hyde, New Arabian Nights*	*Treasure Island, Olalla The ebb-tide*
Jane Austen	*Emma, Lady Susan, Pride and Prejudice*	*Emma, Lady Susan, Pride and Prejudice*
Oscar Wilde	*The Picture of Dorian Gray, Lady Windermere's Fan The soul of a Man under Socialism*	*The Picture of Dorian Gray, The happy Prince Lord Arthur Savile's crime*

Table 2: Data used in the experiments.

Author	English	Spanish
Charlotte Brontë	34	19
Rudyard Kipling	44	28
Lewis Carroll	10	8
Robert Stevenson	22	14
Jane Austen	24	42
Oscar Wilde	12	18

Table 3: Number of texts after splitting.

While the second approach considers the features described in Section 1.1 as cross-language stylometric features, the third one evaluates them in terms of their ability to "survive" despite the noise inserted by the MT.

We use Logistic Regression (LR), Naive Bayes (NB) and k-Nearest Neighbors (kNN) classifiers in our experiments as implemented in the Weka toolkit (Hall et al., 2009)[4] and linear Support Vector Machine (SVM) classifier implementation from LibLinear (Fan et al., 2008). A few experiments with kernel-based SVM were done with LibSVM (Chang and Lin, 2011). For the kNN after trying different k values we finally set k to 3 for our experiments.

We perform classification as leaving one *novel* out. The difference from the leaving-one-out strategy (taking one example as test data, and the rest as training data), is that we take all documents from one novel as test data, and the rest we use for training. This is done in order to prevent a document to be classified according to novel and not author, which is likely especially in case of lexical features.

4. Dataset

The nature of our research requires us to have documents in two different languages. As it is the first attempt at the task, we have chosen to work with English and Spanish, since those two languages contain a sufficient number of resources. Furthermore, due to the cross-lingual nature of our work it is essential to have documents originated by the same author in two different languages X and Y but without the document in language X being the translation of the other in language Y. To the best of our knowledge, there exists no dataset with these characteristics. Thus, we choose to work with authors of the 19th century for whom we can find novels written in English and translated into Spanish.

As opposed to MT which not only introduces errors but also masks the original author, literary translation of prose is believed not to spoil the original style. Shlesinger et al. (2009) in their research on gender identification have found that whatever markers of the translator's style are inserted during the translation, they are much less robust than the style markers of the original author. This legitimizes our choice for using a dataset based on literary translation as the best available alternative to documents directly written by the authors in a different language.

Table 2 presents our dataset. Even though, some novels are presented in both English and Spanish, the classifier is not trained on the translation of the same novel, since we use the leaving-one-novel-out technique described in Section 3. Each novel was split into parts containing 500 sentences. Table 3 shows the number of texts for each author after the splitting.

5. Evaluation

5.1. One-language experiments

We first run the classification for the English and Spanish datasets separately (as a classical authorship attribution task). For both datasets the best accuracy was achieved by linear SVM with bag-of-words features, which was 95% and 93% for English and Spanish respectively. Since the dataset is small, we set the frequency threshold for the lexical features to two. We also tried excluding features with frequencies lower than five, but the accuracy was worse. The accuracy of classification based on high-level features (HLF) for English and Spanish achieved up to 76% and 66% respectively. These results suggest that HLF are quite good as stylometric features for the plain AA task (more than 40% better than random baseline), however do not achieve state-of-the-art performance.

5.2. Cross-language experiments

We performed the experiments on CLAA as described in Section 3. Table 4 presents the experimental results. Surprisingly, the MT approach (the first approach in Section

[4] http://www.cs.waikato.ac.nz/ml/weka/

3) shows relatively good performance: 88% accuracy with bag-of-words features and Naive Bayes, when the same experiment on the one-language data achieved up to 95% accuracy. One of the possible explanations for such high accuracy of the MT method may be in the fact that Google Translate is a statistical machine learning tool, which uses very large amounts of data for training. Since the data we used is freely available online it is likely that these data also appeared among the documents Google Translate was trained on.

The experimental results also allow us to conclude that word n-grams show higher accuracy than character n-grams. Character n-grams may be a less appropriate feature for the MT approach for CLAA, because, as opposed to word-level n-grams, character-level n-grams capture not only lexical but also syntactic information, which is probably too much influenced by the MT.

Pure high-level features, i.e. extracted separately for the two languages (approach 2 in Section 3), lead to poor performance. As the error analysis we perfomed shows, this is due to the difference in the resources used to extract the features. We have manually checked some of the HLF vectors: sentiment and emotional components of the vectors constructed for the same text in English and Spanish are far too different. For example, looking at the HLF vectors constructed for first split of "The Picture of Dorian Gray", we expected the vectors for Spanish and English to be similar. However, we noticed that, for example, the frequency of joy markers for English was almost five times as high as for Spanish. At the same time the number of anger words was higher in the Spanish text. There were similar observations for other novels, however, there is no obvious pattern (e.g. one feature being higher for one language). And this is because the resources we use for their extraction are too different: for instance, words marked as expressing anger in Spanish according to the Spanish Emotion Lexicon are not always translations of those expressing anger in English according to Wordnet-Affect. Thus, HLF extraction for different languages separately requires preliminary resources tuning.

However, when the same features are used jointly with MT (approach 3 in Section 3), and in this case the extraction is done with the same (English) resources, the accuracy is comparable with the one of the MT approach. It is worth noting that the highest accuracy of 88% is achieved when using about 200 000 features, while with only sixteen features described in Section 1.1 we get up to 79% accuracy. This accuracy is similar to one achieved in the one-language setting[5]. It means that the suggested features really "survive" machine translation. Moreover, this accuracy is even higher than in one-language setting for the Spanish data. This allows us to conclude that the resources used to extract the features for the English data are more appropriate to use within CLAA task than those of Spanish.

The experimental results showed that when dealing with the high-level features, kNN classifier performed much better than linear classifiers, so we also run experiments using SVM with different kernels (polynomial with different degrees, RBF kernel and sigmoid kernel). The best accuracy of 61% was achieved when using a quadratic kernel, however this is still worse than the accuracy of kNN.

We also evaluated the approach on pairwise classification, which is a much easier task comparing to the six-class classification described above. In this case the HLF do almost as well as the MT approach. The averaged accuracies are presented in Table 5.

[5]the difference is probably due to the diffence of the datasets: in one-language experiments the English dataset is used for testing and in cross-language experiments the testing is done on Spanish dataset translated to English.

Features	**LR**	**NB**	**SVM**	**kNN**
MT + word 1,2,3-grams	0.76	**0.88**	0.86	0.18
MT + char 2,3-grams	0.50	0.69	0.66	0.57
MT + POS 2,3-grams	0.45	0.46	0.59	0.36
MT + HLF	0.52	0.12	0.26	**0.79**
Pure HLF	0.28	0.20	0.31	0.31
Random Baseline		0.21		

Table 4: Accuracy of CLAA (6-class classification) performed with various features and classifiers.

Features	**LR**	**NB**	**SVM**	**kNN**
MT + word 1,2,3-grams	**0.97**	0.83	0.78	0.48
MT + char 2,3-grams	0.93	0.77	0.93	0.76
MT + POS 2,3-grams	0.83	0.71	0.69	0.75
MT + HLF	0.78	0.90	0.81	**0.95**
Pure HLF	0.61	0.60	0.65	0.66
Random Baseline		0.57		

Table 5: Accuracy of pairwise classification.

6. Future Work

In this study we have focused on introducing and evaluating the task of CLAA on a set of literary translations. We believe that future research should focus on social media-based ones, which, given the nature of the task, seem to be a suitable testbed. However, we expect that datasets constructed from social media texts will introduce further challenges; the language of the Internet often contains non-standard spelling and punctuation (Eisenstein, 2013), which is hard to be analyzed by current NLP tools.

We hypothesize that the good accuracy of the MT-based approach is achieved due to the high quality of the MT component. Thus, when applying CLAA to low resource languages, we would expect that the quality of the translations would decrease, leading to the decrease in the performance of the MT-based approach. Therefore we also plan to focus on the development of cross-language stylometric features, that avoid the need of translation since they are independent of the target language and can be extracted directly from the original texts. On the one hand, we suppose that the features we propose in Section 1.1 could perform better if the difference between the resources is diminished. On the other hand, other stylometric features used in plain authorship attribution, e.g. function words and vocabulary richness, could be also adapted for cross-language setting.

Finally, we plan to adapt techniques for direct transfer of classifiers to other languages by using as features cross-language word clusters (Täckström et al., 2012) or more sophisticated cross-language distributed word representations (Klementiev et al., 2012).

7. Conclusions

In this work, we have introduced the novel task of cross-language authorship attribution and we have proposed two methods for approaching this task, an MT-based approach and an approach based on cross-language features. While the former achieved promising results, the performance of the latter was rather low due to some weakness of the method that pertain to the quality of the resources for languages other than English. Finally, we showed that the combination of the two methods yield comparable results to the one based only on pure lexical features.

8. Acknowledgments

The work of Dasha Bogdanova was supported by Science Foundation Ireland (Grant 12/CE/I2267) as part of the CNGL Centre for Global Intelligent Content, School of Computing, Dublin City University.

9. References

Atserias, Jordi, Villarejo, Luis, Rigau, German, Agirre, Eneko, Carroll, John, Magnini, Bernardo, and Vossen, Piek. (2004). The MEANING Multilingual Central Repository. In *In Proceedings of the Second International WordNet Conference*.

Chang, Chih-Chung and Lin, Chih-Jen. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Dunn, Rita, Beaudry, Jeffrey, and Klavas, Angela. (1989). Survey of Research on Learning Styles. 46(6).

Eisenstein, Jacob. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia, June. Association for Computational Linguistics.

Esuli, Andrea and Sebastiani, Fabrizio. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation*, LREC '06.

Fan, Rong-En, Chang, Kai-Wei, Hsieh, Cho-Jui, Wang, Xiang-Rui, and Lin, Chih-Jen. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Franco-Salvador, Marc, Gupta, Parth, and Rosso, Paolo. (2013). Cross-Language plagiarism detection using a multilingual semantic network. In *Proceedings of the 35th European conference on Advances in Information Retrieval*, ECIR '13.

Gamon, Michael. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of International Conference of Computational Linguistics*. COLING '04.

Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, and Witten, Ian H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Hirst, Graeme and Feiguina, Olga. (2007). Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, 22(4):405–417.

Keselj, Vlado, Peng, Fuchun, Cercone, Nick, and Thomas, Calvin. (2003). N-gram-based Author Profiles for Authorship Attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics*, PACLING '03.

Klein, Dan and Manning, Christopher D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03.

Klementiev, Alexandre, Titov, Ivan, and Bhattarai, Binod. (2012). Inducing Crosslingual Distributed Representations of Words. In *Proceedings of the International Conference on Computational Linguistics*, COLING '12.

Navigli, Roberto and Ponzetto, Simone Paolo. (2010). BabelNet: building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10.

Panicheva, Polina, Cardiff, John, and Rosso, Paolo. (2010). Personal Sense and Idiolect: Combining Authorship Attribution and Opinion Analysis. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC '10.

Peng, Fuchun, Schuurmans, Dale, Keselj, Vlado, and Wang, Shaojun. (2003). Automated Authorship Attribution with Character Level Language Models. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '03.

Perez-Rosas, Veronica, Banea, Carmen, and Mihalcea, Rada. (2012). Learning Sentiment Lexicons in Spanish. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*.

Petrov, Slav, Das, Dipanjan, and McDonald, Ryan. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Potthast, Martin, Barrón-Cedeño, Alberto, Stein, Benno, and Rosso, Paolo. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation*, 45(1):45–62.

Shi, Lei, Mihalcea, Rada, and Tian, Mingjun. (2010). Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10.

Shlesinger, Miriam, Koppel, Moshe, Ordan, Noam, and Malkiel, Brenda. (2009). Markers of translator gender: do they really matter? 38:183–198.

Sidorov, Grigori, Miranda-Jiménez, Sabino, Viveros-Jiménez, Francisco, Gelbukh, Alexander, Castro-Sánchez, Noé, Velásquez, Francisco, Díaz-Rangel, Ismael, Suárez-Guerra, Sergio, Treviño, Alejandro, and

Gordon, Juan. (2012). Empirical Study of Opinion Mining in Spanish Tweets. In *LNAI 7629-7630*.

Stamatatos, Efstathios. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*.

Strapparava, Carlo and Valitutti, Alessandro. (2004). WordNet-Affect: an Affective Extension of WordNet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC' 04.

Täckström, Oscar, McDonald, Ryan, and Uszkoreit, Jakob. (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12.

Toutanova, Kristina, Klein, Dan, Manning, Christopher D., and Singer, Yoram. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03.

Wan, Xiaojun. (2011). Bilingual co-training for sentiment classification of chinese product reviews. *Comput. Linguist.*, 37(3):587–616.