Cross-language authorship attribution of short online texts

s4598350
Radboud University, Nijmegen, The Netherlands
n.wessel@student.ru.nl

ABSTRACT

Authorship attribution is a field that gained much interest in recent years. However, the field has mostly focussed on monolingual data sets. It is therefore interesting to see how authorship attribution performs on data sets that contain mixed languages. In this paper we explore this on a Dutch-English data set crawled from the social media website Reddit. We show that with simple n gram and POS features, it can indeed be helpful to combine data sets into a bilingual set when individual monolingual sets are small. We also find that the chosen approach does not perform well on a training set that is one language and test set that is another.

1 INTRODUCTION

The idea that an author has a specific writing style is almost as old as writing itself. Stylometry, the technique of using writing style to identify an author or characteristics of an author, has been used for hundreds of years, but has seen a boom of interest with the rise of machine learning techniques. Much research has been done in the last couple of years into authorship attribution (i.e. identifying an author by style), and how to improve it. However, there has been little interest so far in authorship attribution across languages. The current paper tries to do some preliminary, exploratory work in this field. We do this by using a bilingual Dutch English data set from Reddit. We look at the accuracy scores of one language, and of combinations of the two languages.

In this paper, we are interested in if it is possible to perform authorship attribution across languages, and if so, how far can we go and what kind of features play a roll. Our test cases are motivated by real world use cases. For example, it might be possible that one has two rather small data sets in different languages. It is then useful to know whether combining the two data sets is a good idea or not. It might also be the case that one has a classifier trained in one language but a set that needs to be classified in another language. Our test cases look into these problems.

In the rest of this paper, we will describe our results. The next section briefly discusses related work. After that, the used approach is explained. Then, the results will be discussed. Lastly, we will look at the implications of this paper.

2 RELATED WORK

Authorship attribution is based on the idea that writers have a distinctive style that remains the same over texts. In most modern day authorship attribution tasks, one has a list of documents and a list of possible authors. The algorithm tries to extract telling features from the texts in the training set that can be used to classify previously unseen texts. Authorship attribution can therefore be seen as a classification problem, where the classes are the different authors.

There are two types of approaches to authorship attribution. In one approach, the algorithm builds a profile and then tries to match a new document to a profile. In the other approach the algorithm retains the original documents, and then tries to classify a new document by finding similar documents.

For possible features, many things have been proposed, such as N-grams [5], function words [3], POS n-grams [3] and syntactic features obtained by parsing [2]

The question now is whether this style also remains the same over different languages. Many of the features mentioned before do not transfer directly over to a new language. N-grams, for example, are monolingual.

Some preliminary research into style across language has found that there are many features that do transfer. It has been shown that, for example, the length of words transfers across languages: if a person is likely to use long words in one language, they are also likely to use long words in another language. The same holds for the use of hashtags on Twitter [4].

[1, 7] have shown that cross-language authorship attribution is possible and can actually perform very well on little data. However, they used lengthy documents (books) for their application, while we are interested in small online posts. While [7] used non-language specific features, [1] used machine translation to classify literary works. They showed that machine translations can provide pretty high accuracy, especially combined with other features.

3 APPROACH

As mentioned before, we are particularly interested in cross-language authorship attribution on smaller snippets of text. In order to easily find bilingual data, we have chosen to crawl from a social media website called Reddit[1]. On Reddit, users can post after making an account. Most accounts are pretty much anonymous in that the user does not use their real name. Reddit is divided into several fora, called subreddits, each with their own theme or common shared interest. While the main language of Reddit is English, some subreddits have one or multiple other languages that users are allowed to post in. This provides us with the right opportunity to gather bilingual data.

In order to gather the data, we crawled `r/thenetherlands`, a subreddit dedicated to The Netherlands. The data used for this research project was crawled from Reddit directly via its json api[2]. The subreddit is Dutch English bilingual with a preference for Dutch posts (i.e. if the user speaks both, they are encouraged to post in Dutch). We used the main page of this subreddit as a starting point for our crawl. The main motivation for this is that since most of the posts are Dutch, the users are expected to speak Dutch. Because

TxMM'19-'20, Radboud University, Nijmegen, Netherlands

[1] https://www.reddit.com
[2] https://www.reddit.com/dev/api/

they use Reddit, they probably also speak English; the website itself (i.e. the skeleton) is only available in English.

We crawled the first few pages of `r/thenetherlands` and made a list of all users we found. We then crawled their Reddit profiles for all their posts. We only took 'self' posts, meaning that the post is a text post and not a link to a video or website. Links barely contain any text written by the user except for the title. We only included 'self' posts that consisted of at least 250 characters. In order to determine the langauge of a post, the Python library Polyglot was used[3]. This library determines which language(s) occur in a post and with how much certainty it can say so. Posts that had a less than 90% certainty for the most often occurring language, or had any other language than Dutch or English as the most often occurring language, were removed from the data set, as we are interested in only almost fully English or Dutch posts.

At this stage, we had 73 authors. We then removed the authors that had less than 25 posts in either Dutch or English. This left us with 21 authors. At that point, we had 1409 documents in Dutch and 2971 in English.

As the Dutch document list is significantly shorter than the English document list, in the subsequent classify phase, we made sure to use the same amount of texts for training and testing each time, i.e. 60% of 1409, so 845 for training, and 564 for testing. This ensures that the different conditions can be compared. Previous research has shown that more data usually gives better results [6]. Therefore it is necessary to keep the training set in similar size in terms of documents and different authors. The test set was also kept the same size in order to obtain the same level of accuracy on the accuracy score.

3.1 Features

For this project, we mostly relied on the N-bag of words method. We used both 1-bags and 2-bags, as they have been shown to be useful in authorship attribution [5], also when using machine translation [1]. Besides making N-bags of the original words, we also made POS N-bags. Besides these features, we used a last category of features that were proven useful in previous research on (online) authorship attribution [8]:

- The number of new lines
- The number of spaces
- The number of two subsequent spaces
- The number of words ending in 'ing'
- The number of upper case letters
- The number of single and double quotes
- The average length of words

All of these features were relativized to the length of the post.

In the rest of the paper, the features are organized in the following sets: 1 bag, 2 bag, 1 POS, 2 POS and other, where other is the list of features given above.

3.2 Testing conditions

In order to analyze which of the use cases mentioned in the introduction are possible with our setup, we trained and tested the data on several different conditions.

- Dutch texts: these are the original Dutch posts. This set is included to set a baseline for Dutch posts.
- English texts: these are the original English posts by the same authors. This condition is included to set a baseline for English texts.
- The translated set (D→E): this is the set of Dutch posts from the first condition, translated with the Google Translate api[4]. This set is included to see how the automatic translation influences our results.
- The English texts and the translated texts (D→E): this set contained a mix of the original English posts and the translated posts that were originally Dutch. This is one of the most interesting sets, as it directly relates to the first use case, where little data is available in the separate languages, but we could combine the data for a mixed language set.
- The English texts as training set to classify translated texts (D→E): whereas in the previous scenario, we mixed the two different parts, here we keep them separate and use one to classify the other. This is also included because it pertains to the use case of having a classifier for one language and applying it to a data set in an other language.
- The translated texts (D→E) as training and the original English texts as test: this is the same as the previous condition, but the languages switched around.

Testing was done by calculating the features and then training a support vector machine (SVM) from the sklearn library [5]. The maximum amount of iterations was set to 2000. If the algorithm had not converged by then, the state in which it was at that moment was taken. This was done to keep running times relatively low.

The algorithm was run five times on different training and test configurations. In the end, the accuracy score for the five runs was taken and averaged to provide the score for that condition. These scores are reported in the results section.

3.3 Effect of different features

In order to see what kind of features have what kind of effect, we carried out an analysis in which we left one set of features out of the feature set and then calculated accuracy again. We also looked at what removing the normal N-grams or POS N-grams completely would do. These results are also reported in the results section.

3.4 Ethical aspects

While the data on Reddit is publicly available and crawable by anyone, even if one does not have an account, it is the case that users of Reddit are not aware that their data has been used for research. The possibility of this happening is explained in the terms of use of Reddit, so we think it is acceptable to conduct this type of research.

4 RESULTS

In this section, we will first discuss the results obtained on the whole feature set, and then walk through the results obtained by leaving out feature groups.

[3] https://pypi.org/project/polyglot/

[4] https://pypi.org/project/googletrans/

[5] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

Table 1: Scores in the different conditions

Condition	Score
Dutch	0.5667
Dutch translated	0.5074
English	0.6028
Mixed	0.4940
E training, D classify	0.1915
D training, E classify	0.0511

4.1 Complete feature set

The results for the complete feature set as described in the previous section can be found in table 1. One can see that the normal English text scores the best of all. The original Dutch text is a close second. After that, the Dutch translated text and mixed text score pretty similar. This suggests that it might be feasible to combine bilingual language sets and classify them together. It seems that most of the accuracy is lost by translating the Dutch posts to English. Machine translation apparently removes a bit of a person's personal writing style. One possible explanation for this is that the machine translation chooses more general translations, and that therefore certain characteristics are lost.

In order to see whether it would be better to use a mixed set or two smaller sets separately, we also ran the algorithm on half the set sizes. This gave an accuracy of 0.5063 for Dutch and 0.6001 for English. This shows that classifying Dutch actually benefits from joining the two data sets, while for English it does not matter as far as we can tell from our limited setup.

We can also see that both the conditions where we use one language to classify the other score particularly poorly. It is interesting to see that the one way works much better than the other way. It looks like using one language to classify another is much harder than mixing the data. This can be expected, since with mixing the data, the classifier has at least seen parts of both data sets, while in the other case, one data set is completely new.

4.2 Partial feature set

The results for the partial feature sets are shown in figure 1. First of all, we see that most scores are quite close to each other, except for the score that is obtained when removing all normal N-grams from the feature set. This causes quite a big drop in accuracy in all conditions except the last one. This is in line with previous research finding that N-grams make an useful feature set for authorship attribution. We can also see that this is still the case with translated texts, although less so; the difference in scores is less in the Dutch translated condition, than in the original Dutch condition. It seems that normal N-grams are provide more information than POS N-grams, and this is persistent across conditions.

The removal of all other features on the other hand, barely makes a difference in the accuracies. This is perhaps a bit surprising. We see that removing the POS 1-grams or POS 2-grams can actually improve the accuracy in some cases. This might be because of redundancy compared to each other. We also see that removing the POS altogether barely has an impact on accuracy, even for English, of which the POSer is known to be pretty good.

5 DISCUSSION AND OUTLOOK

The results raise several questions. For one, why English training-Dutch classifying gives better results than the other way around. A possible explanation is that Google Translate generalizes the posts, and that they are therefore hard to train on for the classifier. Google Translate picks the most common translation for a text, after all. The English posts retain their style and therefore can train the classifier. Then the Dutch translated posts might just retain enough features to classify them somewhat, but not sufficient amount of features to really train the classifier. However, this is all speculation and cannot be known for certain. It could be an interesting future project to look further into this phenomenon.

We also saw that POSing often barely affects accuracy, while previous research has shown that it should. It might be the case that POSing does not contain much new information compared to the traditional N-grams. It would be interesting for future researchers to replicate this research and see whether they obtain similar results.

Looking at the cases that remain difficult in most, if not all conditions, we see some general problematic cases. For example, posts that mostly consist of a link and little text are difficult to classify, as multiple users make those. When crawling, our selection criteria was the amount of characters of a post and not the amount of words. Therefore, these link posts were often added to the data set. Accuracy could probably be improved by removing those from the data set. However, if one does that, one should think about whether the data set then still is representative for the goal.

5.1 Limitations

While the results are promising, there are several limitations of this project that should be kept in mind. First of all, the languages chosen for this research paper are very similar and therefore provide for easy translations. For example, as both languages use determinate articles ('the' in English and 'de' and 'het' in Dutch), in many of the same situations[6], we can assume that counting the amount of times they were used in one language will transfer reasonably well to the other language. However, features of languages that use determinate articles in completely different ways, or languages that have no determinate articles such as Chinese would not transfer well at all. This is a severe limitation in the chosen method of this project, namely bag of words with word for word translations. It remains to be seen how this method fares under more different languages.

We also only selected tweets of a certain length, and only included authors with a certain number of posts. This means that our results will not generalize as well to other users with less posts or less lengthy posts. This is a more general problem of authorship attribution that has been mentioned in the literature [6].

Another limitation is that we might be implicitly using non-stylistic information, such as topic information. Reddit users might specifically only post in certain domains so that their language besides stylistic information is also recognizable because of keywords of their favorite topics. We tried to minimize topic preference by sampling the users from the general subreddit r/thenetherlands,

[6] One should note that there are some slight differences in the exact use of the articles. For example, Dutch says 'in het Nederlands', while English would say 'in Dutch'. However, these are relatively few cases.

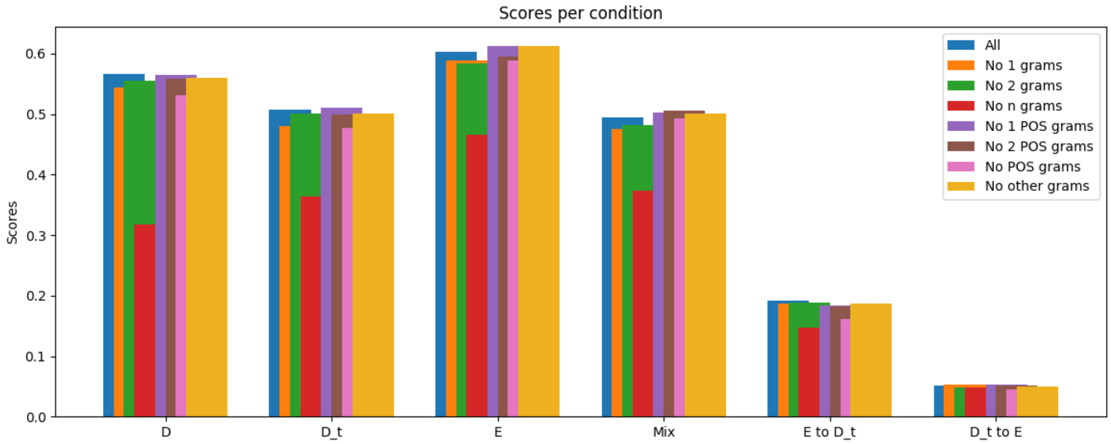

Figure 1: Accuracy scores for each of the conditions with different sets of features. D = Dutch, D_t = Dutch translated to English, E = English, Mix = mixed English - Dutch translated, E to D_t = English as train and Dutch as test set, D_t to E = Dutch as training and English as test set.

which has no real specific topic as long as it has to do with the country (i.e. people share political views, life experiences and many more things there). However, outside that subreddit users could mostly frequent one other sub, which might make identifying those users easier. There were also some double posts in the data set, as some users posted the same post multiple times. These things will always be the case when identifying Reddit users, so is not a limitation in that context, but it remains to be seen if the results also generalize to other online platforms.

5.2 Concluding

In short, we have looked into authorship attribution of short online posts across two languages. The results show that mixing two languages can actually be useful, but that the chosen approach does not work well for a classifier in one language applied to another language. This opens up interesting avenues for further research.

ACKNOWLEDGMENTS

Thanks to dr. I.H.E. Hendrickx, prof. dr. M.A. Larson, K. Laus and J. Winkler for their advice on this project. Special thanks to my parents for proofreading.

REFERENCES

[1] Dasha Bogdanova and Angeliki Lazaridou. 2014. Cross-Language Authorship Attribution.. In *LREC*. Citeseer, 2015–2020.

[2] Olga Feiguina and Graeme Hirst. 2007. Authorship attribution for small texts: Literary and forensic experiments. *Literary and linguistic Computing* 22, 4 (2007), 405–417.

[3] Michael Gamon. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 611.

[4] Patrick Juola, George K Mikros, and Sean Vinsick. 2019. Correlations and potential cross-linguistic indicators of writing style. *Journal of Quantitative Linguistics* 26, 2 (2019), 146–171.

[5] Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, Vol. 3. sn, 255–264.

[6] Kim Luyckx and Walter Daelemans. 2011. The effect of author set size and data size in authorship attribution. *Literary and linguistic Computing* 26, 1 (2011), 35–55.

[7] Raheem Sarwar, Qing Li, Thanawin Rakthanmanon, and Sarana Nutanong. 2018. A scalable framework for cross-lingual authorship identification. *Information Sciences* 465 (2018), 323–339.

[8] Thamar Solorio, Ragib Hasan, and Mainul Mizan. 2013. A case study of sockpuppet detection in Wikipedia. In *Proceedings of the Workshop on Language Analysis in Social Media*. 59–68.