Semi-Supervised Priors for Microblog Language Identification
Carter, S.C.; Tsagkias, E.; Weerkamp, W.

Published in:
DIR 2011: Dutch_Belgian Information Retrieval Workshop Amsterdam

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
ABSTRACT

Offering access to information in microblog posts requires successful language identification. Language identification on sparse and noisy data can be challenging. In this paper we explore the performance of a state-of-the-art n-gram-based language identifier, and we introduce two semi-supervised priors to enhance performance at microblog post level: (i) blogger-based prior, using previous posts by the same blogger, and (ii) link-based prior, using the pages linked to from the post. We test our models on five languages (Dutch, English, French, German, and Spanish), and a set of 1,000 tweets per language. Results show that our priors improve accuracy, but that there is still room for improvement.

Categories and Subject Descriptors
H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing

General Terms
Algorithms, Theory, Experimentation, Measurement

Keywords
Language identification, microblogs, semi-supervised priors

1. INTRODUCTION

Microblogging platforms such as Twitter have become important real-time information resources [4], with a broad range of uses and applications, including event detection [8, 10], media analysis [1], and mining consumer and political opinions [6, 9]. Microbloggers participate from all around the world contributing content, usually, in their own native language. Language plurality can potentially affect the outcomes of content analysis, and we therefore aim for a monolingual content set for analysis. To facilitate this, language identification becomes an important and integrated part of content analysis. In this work, we address the task of language identification in microblog posts.

Language identification has been studied in the past (see Section 2 for previous work in this field), showing successful results on structured and edited documents. Here, we focus on an other type of documents: user generated content, in the form of microblog posts. Microblog posts (“tweets,” “status updates,” etc.) are a special type of user generated content, mainly due to their limited size, which has interesting effects. People, for example, use word abbreviations or change word spelling so their message can fit in the allotted space, giving rise to a rather idiomatic language that is difficult to match with statistics from external corpora.

To address this effect, we use language models trained on microblog posts. To account for very short ambiguous (in terms of what language) microblog posts, we go a step further and introduce two semi-supervised priors, and explore the effects on accuracy of (i) a blogger-based prior, using previous microblog posts by the same blogger, and (ii) a link-based prior, using content from the web page hyperlinks within the post.

In particular, we aim at answering the following research questions: (i) What is the performance of state-of-the-art language identification for microblogs posts? (ii) What is the effect on identification accuracy of using language models trained on microblog posts? (iii) What is the effect on accuracy of using blogger-based and link-based priors? This paper makes several contributions: (i) it explores the performance of state-of-the-art language identification on microblog posts, (ii) it proposes a method to help identification accuracy in sparse and noisy data, and (iii) it makes available a dataset of microblog posts in for others to experiment.

The remainder of the paper is organized as follows: in Section 2 we explore previous work in this area. In Section 3 we introduce our baseline model, and the semi-supervised priors. We test our models using the setup detailed in Section 4, and in Section 5 we present and analyze the results. Finally, we conclude in Section 6.
We consider three types of language models for identifying a document’s language. More precisely, we use the TextCat algorithm for language identification on our microblog post set and study the effect on TextCat accuracy of language models trained on different data sets. We consider three types of language models for: (i) out-of-the-box, which uses the training data supplied by TextCat and we set this as our baseline, (ii) microblog, for which we use a training set of posts from our target platform to re-train TextCat, and (iii) combined, that merges n-grams from both other models.

Let $n$ be the total number of languages for which we have trained language models and $i \in \{1, \ldots, n\}$ denote the corresponding model for a language. For each post $p$ we define a language vector

$$\lambda_p = (\lambda_{p,1}, \lambda_{p,2}, \ldots, \lambda_{p,n})$$

(1)

where $\lambda_{p,i}$ is a score denoting the distance between $p$ and language $i$ (the smaller the distance the more likely is $p$ to be written in language $i$). TextCat scores are not normalized by default and therefore we define $\hat{\lambda}_p$ using the z-scores: $\hat{\lambda}_p = (\lambda_{p,1}, \lambda_{p,2}, \ldots, \lambda_{p,n})$. We call vectors constructed from the microblog post itself content-based identification vectors and for post $p$ we write $\hat{\lambda}_p$.

### 3.1 Semi-supervised priors

On top of the language identification on the actual post, we use two semi-supervised priors to overcome problems due to sparseness or noise. Our priors are (i) semi-supervised, because they exploit classifications of the supervised language identifier on unlabeled data, for which we do not know beforehand the true language, to improve the accuracy of our baseline classifiers, and (ii) priors, because they allow us to identify the language of a post without the content-based identification. We propose the use of two priors:

**Blogger-based prior**: behind each post is a blogger who wrote it, and probably the current post is not her first; there is a post history for each blogger the content of which can be beneficial for our purposes. By identifying (or guessing) the language for previous posts by the same blogger, we construct a blogger-based prior for the current post.

Let $P = \{p_1, \ldots, p_k\}$ be a set of posts predating $p$ from blogger $u$. For each $p_i \in P$, we use the microblog language models, and construct $\lambda_{p_i}$, as explained before. We then derive a blogger-prior from the average of content-based identification vectors of previous posts:

$$\mu_{p} = \frac{1}{|P|} \sum_{i \in P} \hat{\lambda}_{p_i},$$

(2)

**Link-based prior**: posts in microblogs often contain features like links or tags. Links refer to content elsewhere on the web, and this content is often of longer text length that the post itself. We identify the language of the linked web page, and use this as link-based prior for the post that contains the link.

Let $L = \{l_1, \ldots, l_i\}$ be a set of links found in post $p$. For each web page $l_i \in L$ we apply the out-of-the-box model to its content, and construct a link-based prior vector from the average of content-based identification vectors of web pages found in $p$:

$$\nu_{p} = \frac{1}{|L|} \sum_{i \in L} \hat{\lambda}_{l_i}. $$

(3)

Having constructed three vectors (content, blogger and link-based) with scores for each language, we combine the three vectors using a weighted linear combination. More formally, we identify the most probable language for post $p$ as follows:

$$\text{lang}(p) = \arg\min_{v \in \{C, B, L\}} \frac{1}{|v|} \sum_{i \in v} w_i \hat{\lambda}_p, $$

(4)

where $v = \{C, B, L\}$ and $\sum_{i \in v} w_i = 1$. Finally, language $\lambda$ that is closest to the language profile (i.e., has the lowest score) is selected as language for post $p$.

### 4. EXPERIMENTAL SETUP

For testing our models we need a collection of microblog posts. We collect these posts from one particular microblog platform, Twitter. We test our models on a set of five languages, Dutch, English, French, German, and Spanish, and gather an initial set of tweets (Twitter posts) by selecting tweets on their location. From this initial sample, we manually select 1,000 tweets in the appropriate language.

In case of a multilingual tweet, we assign the language that is most “content-bearing” for that post. For training purposes, we split each set in a training set of 500 tweets and a test set of 500 tweets. We construct test and training sets by taking one every other tweet so both sets contain approximately the same language.

TextCat allows us to select the number of n-grams we want to use for profiling our language and documents. Preliminary experimentation with this parameter revealed that the standard value (top 400 n-grams) works best, and we use this value for the remainder of the experiments. In our experiments we use fixed weights for the three language vectors; our intuition is that the content-based identification should be leading, supported by the blogger-based prior. Since people can link to pages in other languages as well, we assign least weight to the link-based prior. The actual weights are given in Table 2.

<table>
<thead>
<tr>
<th>Run</th>
<th>$w_C$</th>
<th>$w_B$</th>
<th>$w_L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>microblog + blogger-based</td>
<td>0.66</td>
<td>0.33</td>
<td>-</td>
</tr>
<tr>
<td>microblog + link-based</td>
<td>0.75</td>
<td>-</td>
<td>0.25</td>
</tr>
<tr>
<td>microblog + both priors</td>
<td>0.50</td>
<td>0.33</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Table 2: Weights for runs, results are shown in Table 3.

We report on accuracy (the percentage of tweets for which the language is identified correctly) for each language, and overall. In total we look at six runs: the out-of-the-box language model, the

1http://www.let.rug.nl/~vannoord/TextCat/

3The actual dataset will be made available online
microblog language model, the combined language model, the microblog model with each prior separately, and the microblog model with both priors.

5. RESULTS AND ANALYSIS

In Table 3 we present the accuracy of our runs for all languages. The results show that language identification on short posts in microblogs is not as straightforward as it is on longer pieces of text. Training the n-gram-based approach on the target corpus obviously gives much better results, but accuracy is still limited. Incorporating the semi-supervised priors does lead to an increase in accuracy for all languages, and especially the combination of the blogger-based and link-based priors outperforms other approaches.

<table>
<thead>
<tr>
<th>Run</th>
<th>Dutch</th>
<th>English</th>
<th>French</th>
<th>German</th>
<th>Spanish</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content-based identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Out-of-the-box</td>
<td>90.6%</td>
<td>85.0%</td>
<td>86.0%</td>
<td>93.6%</td>
<td>82.2%</td>
<td>87.5%</td>
</tr>
<tr>
<td>Microblog</td>
<td>90.4%</td>
<td>91.6%</td>
<td>92.2%</td>
<td>95.4%</td>
<td>85.2%</td>
<td>91.0%</td>
</tr>
<tr>
<td>Combined</td>
<td>92.2%</td>
<td>89.0%</td>
<td>91.6%</td>
<td>92.2%</td>
<td>83.2%</td>
<td>89.6%</td>
</tr>
<tr>
<td>Microblog content-based identification + priors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blogger-based</td>
<td>94.6%</td>
<td>93.6%</td>
<td>94.8%</td>
<td>96.4%</td>
<td>84.6%</td>
<td>92.8%</td>
</tr>
<tr>
<td>Link-based</td>
<td>92.0%</td>
<td>90.6%</td>
<td>92.6%</td>
<td>92.8%</td>
<td>83.0%</td>
<td>90.2%</td>
</tr>
<tr>
<td>Both priors</td>
<td>94.4%</td>
<td>95.0%</td>
<td>94.0%</td>
<td>97.2%</td>
<td>85.4%</td>
<td>93.2%</td>
</tr>
</tbody>
</table>

Table 3: Results for baseline content-based identification runs and the combination with the priors.

We notice differences in accuracy between languages: for German, English, French, and Dutch, accuracy is high (although there is room for improvement), for Spanish accuracy is quite low. In the next section we briefly touch on this with some examples of errors made in the identification process.

5.1 Error analysis

In analyzing the posts misclassified by our final classifier using all priors, we group them into four distinct categories: fluent multilingual posts, those containing named entities, automatically generated, and language ambiguous. We give examples in Table 1, and explain each type of error in turn.

Fluent multilingual posts: These are posts which are grammatical sentence with words written in two or more languages. Usually these take the form of a sentence split into two, with both halves in different languages.

Named entity errors: These posts are misclassified because they contain a reference to a foreign language named entity, such as a company or product name, song title, etc. The named entities contained in the post outweigh the correct language tokens in the post in scoring, leading to the misclassification.

Automatically generated posts: These posts are automatically generated by external applications and software, which insert phrases into the post foreign to the language of the user.

Language ambiguous: These posts are misclassified because they only contain a few tokens which could belong to a number of different languages.

6. CONCLUSION

In this paper we explore the performance of an n-gram-based approach to language identification on microblog posts. Given the short nature of the posts, the rather idiomatic language in these (due to abbreviations, spelling variants, etc.), and mixed language usage, we expect language identification to be a difficult task. To overcome the challenges of microblogs, we introduce two semi-supervised priors: (i) a blogger-based prior, using the previous posts of a blogger, and (ii) a link-based prior, using the pages a post links to. Results show that accuracy for 3 out of 5 languages is the best using both priors, and the remaining 2 languages benefit most from the blogger-based prior alone.

Analysis reveals four main categories of errors: fluent multilingual posts, named entity errors, automatically generated posts, and language ambiguous posts. All of these types of errors could, in principle, be overcome using different relative weighting of the priors to the content-based identification.

Although accuracy for most languages is high, we feel that there is room for improvement. Microblogs (and possibly other social media as well) offer several other priors that we have not yet discussed or explored. Bloggers often write posts in reply to a previous post by another blogger: we can take use the language profile of this other blogger as a prior on the current post, e.g., as...
a reply-based prior. In the current setup we did not use tags attached to posts (besides keeping them for identification purposes); a future direction could involve collecting posts with the same tag, and construct a language profile for this tag. We can then use this score as a tag-based prior for language identification. Finally, in our experiments we used fixed weights for combining priors and content-based identification, but we are interested in investigating how weights affect accuracy. We believe weights should be dependent on the individual post: when content-based identification results are close for multiple languages, we might want to lower its weight, and rely more on our priors. Future work aims at finding a proper way of estimating these post-dependent weights.

Acknowledgements
Tsagkias is supported by the Netherlands Organisation for Scientific Research (NWO) under project number 612.061.815.

References