Unsupervised Clustering for Language Identification

Karen Shiells and Peter Pham

December 8, 2010

1 Introduction and Prior Work

The current state of the art in language identification comes from n-gram language models. While these can reach 99% accuracy (Hammarstrom, 2007), they have three major shortcomings. First, n-gram language models are supervised. They require substantial labeled training data in each language in order to be functional. For best results, this training data should also be in the same genre as the text ultimately to be identified. Second, n-grams are most reliable in, and the 99% accuracy cited above comes from, identification of longer texts. Other papers evaluating n-gram language models on shorter texts, specifically 6- to 7-word documents, found accuracies closer to 60% (Vatanen, 2010). Finally, as access to the internet expands across the globe, members of more minority language groups are beginning to produce content. Languages are appearing on the internet for which training data is limited, often noisy, and not freely available in the form of a corpus. In order to get away from these issues with supervised language identification, for our project we will explore unsupervised language identification in a domain where all of the above problems for supervised identification become relevant.

We chose to use Twitter as the source of data for our project because tweets are short, exist in a very large number of languages, and are very unlike any labeled corpus of training data. Twitter imposes a hard limit on each tweet of 140 characters, at which length word n-gram models will almost certainly have difficulties. The language used on Twitter is unlike most corpora in that it includes a large amount of netspeak This includes abbreviations such as u and lol, symbols like :), and intentional misspellings, for example when a user expresses that she is huuuuungry. In addition to these intentional differences from standard written language, users often write from phones or type quickly, leading to a higher rate of typos.

All of these examples are drawn from English, but English is only one of many languages used on Twitter. And while English is unsurprisingly most common, and Portuguese in second place is also a major language with plenty of available data, the third most commonly used language on Twitter is Indonesian. Already at the third most-used language we find one that is not included in standard multilingual corpora, and of course the underrepresentation gets worse as we continue.

While we could potentially solve all of these problems by paying annotators to label a new corpus of Twitter language and training classifiers from that new corpus, this approach requires that annotators re-label whenever a new language community makes its presence known and does not extend to other domains. Instead, we propose an unsupervised approach, which may not reach quite the same classification accuracy, but will instantly discover new languages, can be quickly adapted to new domains, and does not require human annotation.

Previous work on unsupervised language identification has, like most language identification literature, focused on longer documents. The state of the art in this area is Chinese Whispers, an algorithm which clusters the word co-occurrence graph to identify languages. According to the paper in which it was introduced, Chinese Whispers achieved an F1 score over .99 for separating a 7-lingual corpus. These results are comparable to supervised results without need for a labeled corpus. Much like supervised identification, however, the performance of the system decreases significantly when applied to the short, noisy data found on Twitter. For our project, we will attempt to improve the performance of Chinese Whispers on Twitter language identification.

2 Data

In this project, we computed the coocurrence graph for tokens over a set of one million tweets uniformly sampled from the twitter data in (Yang, 2011). We decided to restrict our working set of languages to those that can be approximately tokenized by whitespace because tokenization is not the focus of our project. We used a simple custom tokenizer that discards tweets from non-space separated languages such as Chinese and Japanese. The final coocurrence graph we used contains a total of 300,000 tokens. We did not attempt to normalize the tokens in any way other than making them lowercase. Just as with tokenization, robust normalization would be beyond the scope of our project.

In order to construct the data set on which we would evaluate our objective function, we used an existing supervised language identification API provided by Microsoft Translator. Although this service does not have perfect identification accuracy (nothing would aside from a set of people who speak all of the languages in the corpus), it is good enough to provide the basis for a measure of our unsupervised performance. The test set that we use primarily is sampled uniformly from the whole Twitter data set. However, we also performed some testing on a data set with a balanced number of tweets for each language included.

3 Methods

3.1 Pipeline

Our project has two principal components: initial clustering of tokens (hopefully into languages) and the assignment of an unseen tweet to a cluster. We explored the application of the following concepts to both of these components.

3.2 Purity

The motivation behind our definition of this quantity involves the fact that we do not want nodes appearing at the boundaries of clusters to overly influence the assignment of a tweet to a particular cluster. We define the purity p_i of node i as follows:

$$p_i = \frac{\sum_{n \in N(i)} I_{A(i)}(A(n))}{|N(i)|} \tag{1}$$

Where $N(i)$ is the set of neighbors of the node i and $I_{A(i)}$ is an indicator function that returns 1 if the cluster assignment of node n is the same as that of node i, $A(i)$.

3.3 Authority

As with purity, our definition of authority is movtivated by the properties of our data. In language some words are used with much higher frequency than most words; the usage of words exhibits power law characteristics. Therefore, a word with high frequency will be more indicative of a particular language than a word with low frequency. Words with low frequency can potentially just be a word borrowed from another language if the size of the training data is not large enough. We define authority a_i of node i as follows:

$$a_i = \frac{freq(i)}{\max_i freq(i)} \tag{2}$$

Where N and A are defined as above and $freq(i)$ is the frequency of the token i in the corpus.

3.4 Evaluation

The ideal clustering would be such that given a tweet assignment method A, would assign all of the tweets of the same language to the same category. Our objective function F captures this as follows:

$$F(\phi) = \sum_i \sum_j I(A(D_i), A(D_j)) \tag{3}$$

Where ϕ is the assignment being evaluated, D_i is the i^{th} tweet in the test set and $I(d_1, d_2)$ is an indicator function that returns 1 if d_1 and d_2 are different and in the same cluster or the same and in different clusters.

4 Experiment Descriptions

This section provides and overview of the types of experiments we performed and the motivations behind them.

4.1 Chinese Whispers Validation

In order to determine whether Chinese Whispers is a reasonable starting point for improvement, we compared its performance with the that of a mutlilevel graph cut approach implemented in the Graclus library (Dhillon, 2007).

4.2 Chinese Whispers Variants

4.2.1 Asynchronous vs Synchronous

Although the original Chinese Whispers algorithm calls for synchronous updates to the cluster assignments, we tested the possibility that asynchronous updates would work just as well or better. The motivation behind this modification is that if asynchronous udpdates were to work at least as well as the synchronous updates, then it would be possible to parallelize the algorithm, allowing us to work with larger graphs.

4.2.2 Purity and Authority

In this set of experiments, we incorporate our notions of purity and authority into Chinese Whispers. In other words, when computing the updates, we tried a variety of methods of weighting the votes of nodes.

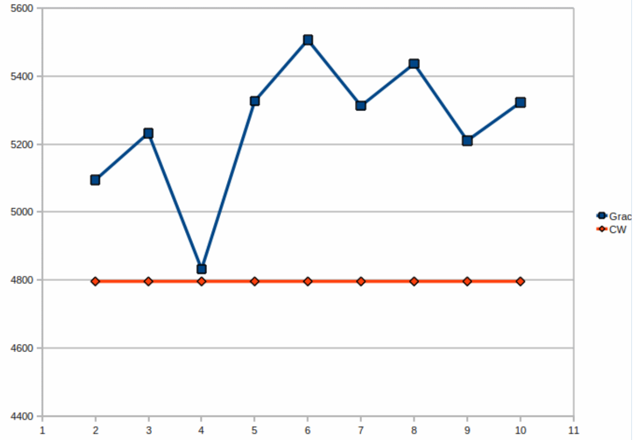
4.3 Cluster Assignment Variants: Purity and Authority

In this series of tests we considered different ways of assigning an unseen tweet to a cluster. This set of tests allowed us to determine whether purity and authority should be applied after or during clustering. More importantly, finding a good weighting scheme is important in the case of tweets because there are so few words in a tweet; it is definitely possible to throw the assignment off with a few words from a different language.

5 Results

5.1 Chinese Whisers: Validation

The results for the first experiment suggest that the Chinese Whispers algorithm is indeed a reasonable starting point. As the following graph shows, the Graclus software provided worse objective values for all values of k, the target number of clusters.



Not only is this the case, but graclus requires apriori knowledge of what the value of k should be. Chinese Whispers performs better even without requiring this knowledge. Furthermore, upon examination of the clusters that Chinese Whispers generates, there seem to be four clusters, which agrees with the fact that setting $k = 4$ yielded the best Graclus clustering.

During the validation process, we also made some general observations about applying Chinese Whispers to Twitter data. We noted that the number of iterations of Chinese Whispers needed to reach the optimal clustering with respect to the objective value is small. Furthermore, we note that the algorithm does not seem to converge with Twitter data as opposed to with the much longer documents used in the original Biemann paper. This suggests that there are not clear divisions between the clusters of the coocurrence graph induced by the Twitter data. In other words, there are many more edges between clusters. One would expect these findings because of the more formal corpus of news articles and legal text used in the original paper. Finally, we would like to note that empirically, the best value of the objective function always comes from the first minimum that is reached during the clustering process.

5.2 Chinese Whispers Variants

5.2.1 Asynchronous vs Synchronous

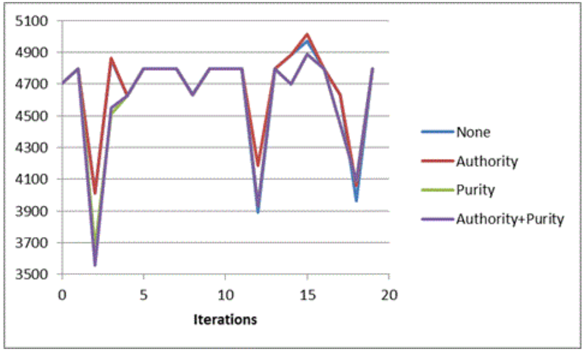
The results for this experiment was disappointing because asychronous updates performs much worse than synchronous updates. Asynchronous updating makes the algorithm converge in only a few iterations, however, on all of the datasets that we tested, the fixed point assigns all nodes the same cluster. This most likely occurs because of the structure of the graph. Once a few nodes change their assignment to a cluster, the rest of the graph quickly switches over to this cluster.

5.2.2 Purity and Authority

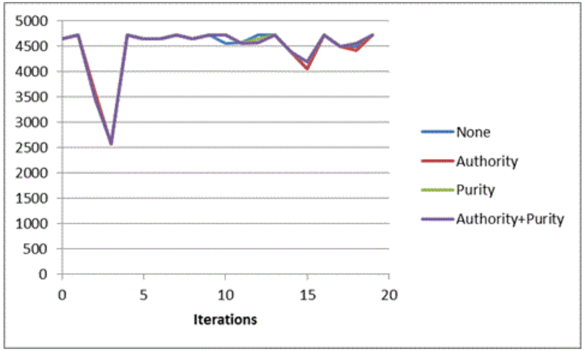
Our experiments with purity and authority show that they are not helpful during the clustering process itself. We tried the following variants of weighting during the clusteing process: purity only, authority only, purity and authority, and the smoothed versions of all three of these (where smoothing involves adding a constant value to weight). The results of applying purity and authority during clustering are similar to those of asynchronity: Within a few iterations, all of the nodes are assigned to the same cluster. Weighting by authority probably leads to a degenerate clustering because the words with the highest weight are more likely to be connected to each other. Thus, they are more likely to influence each other and consequently the whole graph to join the same cluster. Weighting by purity probably leads to degenerate clustering because once a single cluster starts to develop, it will have the greatest weight under this scheme and simply spread over the entire graph.

5.3 Cluster Assignment Variants: Purity and Authority

The following graphs show the effect of different cluster assignment variants.



These results show that our variants do improve the value of the objective function on a dataset that has been sampled from the unbalanced distribution of true Twitter data. The following graph shows the results of clustering on a data set engineered to have a balanced distribution of languages.



Although none of the assignment variants actually improved the objective value, they did not make the objective value worse for any particular clustering. These results suggest that our weighting schemes can only improve cluster assignments in general, but will improve assignments when we are testing data from the training distribution.

6 Conclusion

In our investigation, we found that Chinese Whispers is indeed better suited to the task of language clustering than a graph cut method. We found that changes to the clustering process, including asynchronous update and purity and authority weighting, interfere with the fairly delicate process of inducing multiple clusters on the graph, instead of just one corresponding to the majority language. By applying purity and authority weighting in the cluster assignment, however, we were able to improve our results on a dataset with the natural Twitter distribution of languages. We also tested our variant on a dataset more like the one presented in the original Chinese Whispers paper, which was engineered to provide equal distributions, and found that our improvements were less effective, suggesting that they somehow capitalize on the unevenness of the language distribution, or else cope with it better than the standard Chinese Whispers cluster assignment with uniform cluster assignment. Thus our main contribution from this project is an improved method for assigning short texts to the clusters produced in unsupervised language identification.

7 References

References

[1] C. Biemann and S. Teresniak, *Disentangling from babylonian confusion - unsupervised language identification.* In A. F. Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, 6th International Conference, 2005

[2] C. Biemann, *Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems.* Proceedings of the HLT-NAACL-06 Workshop on Textgraphs-06, New York, USA, 2006

[3] I. Dhillon, Y. Guan, and B, Kulis, *Weighted Graph Cuts without Eigenvectors: A Multilevel Approach.* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 29:11, pages 1944-1957, 2007

[4] Harald Hammarstrom, *A fine-grained model for language identification*. In Proceedings of NEWS-07 Workshop at SIGIR 2007, 2007

[5] Vatanen, Tommi, Jaakko J. Vayrynen, and Sami Virpioja, *Language Identification of Short Text Segments with N-gram Models*. In Proceedings of LREC 2010, 2010

[6] J. Yang, J. Leskovec, *Patterns of Temporal Variation in Online Media*. In Proc. Forth ACM International Conference on Web Search and Data Mining (WSDM '11), 2011