The authors illustrated a new method in this paper to solve the problem of language of Identification which is usually considered as a solved problem. However, as suggested by the author, the performance of previous models always drop on cross-domain data which left space to refine the solution of this task. The idea of maximizing the difference between information gain for languages and that for domains for features selection based on former reliable works is innovative and conforms with the purpose of this work. Several plots and data showing the success in transfer learning make their project actually have a chance to be widely used in the real world. I believe languages are probably uniformly distributed within one dataset according to their description of all datasets, but I cannot find related information about how this model’s performance on the unknown(und) languages. It mentioned that totally 97 languages are trained in this case but obviously there are more languages in the world. Since other languages may not have enough sources to train, it leaves space for them to detect those low resources languages in fewer features.

This paper provided a solution to several tasks including POS tag, chunking, and NER on twitter data. These tasks have several fairly good existing tools but suffer from a significant drop in performance when transferred to twitter data. The author considered several special characters of tweets like lots of infrequent entities, OOV and lack of background information within in short sentences and provided corresponding methods like clustering, T-CAP classifier, and distant supervision. Besides the impressive numbers in this work, I like the idea to consider different tasks separately so that using ‘the most’ suited models for each of them. That can be especially useful for starting in a relatively new domain or dataset. Even I believe the solution mentioned here could work, I still wonder if it can apply to twitter data we would collect today. We know that twitter languages are quickly changed overtime, considering this is a paper in 2011, it can be more convincible if it could survive in the reexamination it today.

In this paper, the authors built probabilistic models to identify AAE dialect based on geoinformation and smartly evaluated by incorporating phonological and syntactic phenomena. Also, an improvement in existing language identification tools is achieved. I believe that capturing the assumption that there’s a correlation between demographics and languages lays a good foundation for their work. While they did not explain why using Gibbs sampling but I think it is a good model in this case since direct sampling is difficult. The contribution of their work is not limited to their insights on the analysis of special characteristics of dialects but with profound implication on social science. I will not surprise that if later works on other dialects of various languages in the world are established in a similar way or more corpus of those languages would be released. While the problem for today or other low-resource languages still exists that how we should explore the geoinformation when lots of tweets of some languages are not geo-enabled.

While this database has already widely used, there are still improving space for it. The first thing is that proper nouns are still limited by the uneven distribution of the domain of corpus. If I search India, the word difference is in the list; If I search Ukraine, the word border is in the list. It’s possible that texts about Ukraine in the dataset always talk about the border problem. More corpus could possibly solve this problem to leverage unevenness.

The second thing is that paraphrase pairs are not always substitutable. For example, if I search for nuts, the result list includes hazelnuts, walnuts, and groundnuts, etc. These specific kinds of nuts cannot replace nuts all the times. This may come from the limitation of this model since the features used in the ranking like n-gram, position, syntactic information can actually be very similar for these words. However, relations among these words are not discovered from that. I think it can be very difficult to find these relations based on the current model. Incorporating distant information from some knowledge bases may help.

Another thing I am wondering is how good this model will be if it applies to a gender-sensitive language. English does not need to deal with that for most cases but for language like French, the database could lose the information about gender if English is set as a foreign language. (Or probably we want the loss of gender information?)