Telling English Tweets Apart: the Case of US, GB, AU

Asmelash Teka Hadgu
L3S Research Center
Hannover, Germany
teka@l3s.de

Netaya Lotze
Westfälische
Wilhelms-Universität
Münster, Germany
lotze@uni-muenster.de

Robert Jäschke
L3S Research Center
Hannover, Germany
jaeschke@l3s.de

ABSTRACT

In this paper, we study how to automatically tell different varieties of English apart on Twitter by taking samples from American (US), British (GB) and Australian (AU) English. We track cities and apply filters to generate ground-truth data. We perform expert evaluation to get a sense of the difficulty of the task. We then cast the problem as a classification task: given a tweet (or a set of tweets from a user) in English, the goal is to automatically identify whether the tweet (or set of tweets) is US, GB or AU English. We perform experiments to compare some linguistic features against simple statistical features and show that character Ngrams are quite effective for the task. Our work is closely related to socio-linguistics, especially research on diatopic varieties, linguistic landscapes, and World Englishes [5].

1. INTRODUCTION

One of the challenges for language identification systems is automatically telling apart similar languages and language varieties. This problem has attracted interest of researchers in recent years. There are dedicated shared tasks: *Discriminating between Similar Languages* (DSL) that took place 2014[1] and 2015[2] targeted at this problem. These shared tasks use excerpts extracted from journalistic texts tagged with the country of origin of the text. In contrast, we investigate the problem in noisy unstructured microposts.

Our work discusses a corpus-based analysis of Twitter messages composed by British, Australian, and American users. The main goal is to detect diatopic varieties of English on Twitter automatically, using a classification algorithm based on linguistic parameters. By doing so, we can address two prominent questions from the fields of socio- and computational linguistics: Is it possible at all, to detect the subtle differences between the varieties of one language, automatically? And if so, which linguistic parameters are the most reliable predictors in an automated classification task. To find these predictors, we have run a small-scale corpus study on varieties of English on Twitter by a group of linguists [6] in which we have annotated dialectal phenomena of approximately 1000 Tweets by hand (N (US) = 320, N (GB) = 320, N (AUS) = 320). This corpus linguistic approach is a more fine-grained approach with a higher ecological validity compared to studies that classify varieties based on what are perceived to be typical differences in spelling conventions in British English vs. other varieties of English (e.g., [7]). Our set of linguistic phenomena that we claim be typical for one or the other variety of English on twitter has been extracted from real tweets of real people and, therefore, is sensitive to the special conversational conditions of computer mediated communication (CMC) [4].

Our contributions are:

- An automatic approach to build datasets for language varieties detection on Twitter
- An expert evaluation to put the difficulty of the task in perspective
- A comparison of linguistic vs. statistical features for classifying language varieties
- By clustering the tweets after geo-location and dialectal linguistic material, we address several questions from the field of socio-linguistics:
  - Can subtle linguistic varieties be differentiated on a very basic lexical and orthographical level (by an algorithm)?
  - How relevant are diatopic varieties for conversational needs in the online environment of a global community?
  - Does the digital landscape of Englishes [2] of the English-speaking twitter blogosphere match the geographical distribution of English varieties on planet earth?

This paper is organized as follows: in Section 2 we review related work on the topic of language variety detection. Then, in Section 3 we present our approach and in Section 4 we outline the setup for our experiments. We present and discuss the results of our experiments in Section 5 and provide a conclusion in Section 6.

2. RELATED WORK

2.1 Gender and Location from Tweets

The problem of inferring location and gender from tweets is highly relevant to our work. Because the binary distinction between 'male' and 'female' language did not match the diversity found on Twitter, [1] ran a cluster analysis to detect Communities of Practice (CoPs) in gender-associated communication styles. From a linguistic perspective, this is interesting because it gives us

[1] http://corporavm.uni-koeln.de/vardial/sharedtask.html

[2] http://ttg.uni-saarland.de/lt4vardial2015/dsl.html

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

a detailed picture of different communication styles used by different CoPs. This is a very modern approach, because they do not put gender stereotypes in fixed categories, but try to relate communication styles with CoPs. Their conclusion is that there are no exact boundaries between genders, as their cluster analysis reveals the underlying heterogeneity.

[3] took a more conservative approach to gender classification of tweets by using pre-determined linguistic features. They distinguish between male and female communication styles and use linguistic indicators for those in their classification task. This approach is closer to ours on varieties of English, but not as close to the communicative reality of CoPs in the anglophone blogosphere as in [1]. A categorical classification by a robust tool does not always capture the linguistic reality adequately.

The two different papers shed some light on the underlying sociological and linguistic challenges we are facing, too. Our approach combines a clearly defined classification task with the sensitivity for the different sub-standard aspects that influence the linguistic structure of English tweets.

2.2 National Dialects and Similar Languages

The closest previous work to our approach is the work by [7] which investigates the task of national dialect identification across Australian, British, and Canadian English. Their approach uses various data sources like national text corpora and webpages crawled from national domains (.au, .ca, .uk), government web portals, and Twitter in a classification scheme. Their findings suggest that there are lexical and syntactic characteristics of these national dialects that are consistent across the different data sources. [7] claim that the classification of language varieties on the tweet level is impossible. In this work we investigate approaches for this task and show that it is solvable.

[9] describes a simple method using word and character n-gram features for classification. He achieves competitive results on the DSL shared task with over 90% accuracy on the test data. By fixing simple file handling and feature extraction bugs, this improved to over 95% which is comparable to the best submitted approaches.

3. APPROACH

3.1 Preliminary Study

The linguistic features that we will use in our approach were identified by expert linguists in a prior study [6]. The initial data of the small-scale study, consisting of 640 tweets from 64 randomly chosen users who self-identified as either British or US-American and collected in 2010, were analyzed using a mixture of automatic part-of-speech and manual tagging. The Australian corpus was built in a parallel way in 2014, using a comparable set of tweets within the same time frame as the original corpus. The small-scale study used a broad approach to analyzing the language found in tweets, therefore the manual tagging encompasses a wide range of categories, including, but not limited to, orthography, typography, lexis, syntax and pragmatics. Particular attention was paid to the features that have been identified as typical of computer-mediated communication, such as non-standard spelling and punctuation as well as the use of emoticons and abbreviations. The small-scale study was part of a larger interdisciplinary project investigating Twitter communication in ten languages [10].

3.2 Features

We explore two groups of features that are commonly used for language identification tasks, namely linguistic and statistical features.

Table 2: Count of users and tweets before filtering

	location	#all tweets	#en tweets	#users
US	Houston	3,096,458	2,655,197	80,815
	New York	11,308,208	9,435,204	168,340
GB	Birmingham	591,394	514,955	38,493
	London	3,762,281	3,061,529	146,664
AU	Perth	81,845	61,710	3,432
	Sydney	224,656	176,886	11,845

Linguistic Features.

These are features selected by linguists to study any systematic difference among varieties of English. We took a subset that are amenable to computation, e.g., use of emphasis, g-dropping, use of emoticons etc.(cf. Table 1). We use the freely available CMU Twitter PoS Tagger [8] to generate PoS tags.

Statistical Features.

We mainly use n-gram features derived from the tweets. These include both character n-grams and word n-grams. We explore a range of values for n from 2 to 6 in a grid search approach to perform the classification.

4. EXPERIMENTAL SETUP

4.1 Dataset

We gathered tweets using the Twitter streaming API[3] by tracking a pair of cities for each of the different varieties of English. We obtained the bounding boxes for the cities from Klokan Technologies.[4] These bounding boxes were used to filter tweets by their geo location from the Twitter API. The crawl duration was three weeks from Apr 7th to Apr 27, 2014, inclusive. Table 2 gives the number of total and English-only tweets crawled by city and country.

4.2 Preprocessing

Not all tweets originating from the cities we tracked are good examples for their respective variety. This can be due to people who had visited these places at the time of crawl or whose mother tongue is not English. Therefore, we apply some filtering to clean the dataset.

4.2.1 User-Level Filtering

We remove users from the dataset based on some properties of their Twitter profiles:

Location. Users can enter location information into their Twitter profile. These user-generated locations can serve as proxies to determine where users come from. We performed a re-crawl of users and their tweets one year after the first snapshot was taken. We then checked whether the user location (as given in their profile) refers to the same country as before and whether it matches to the country from the location where the users' tweets were tracked. Note that the location information in the profile is free-form text. To identify the country from these texts, (i) for user locations with coordinates (i.e., strings such as "iPhone: -32.057101,115.747066" or "ÜT: 41.788018,-72.716657") we use OpenStreetMap Nominatim to resolve their corresponding location ("Perth (AU)" and "New York (US)", respectively), and (ii) for the remaining user locations,

[3] https://dev.twitter.com/streaming/overview

[4] http://boundingbox.klokantech.com/

Table 1: Overview of linguistic features

feature	description	linguistic area
omission of apostrophes	check for omission of apostrophe in contracted verbs without the apostrophe, e.g., dont, wont	orthography
use of lower case i	check if personal pronoun I is written in lower case	orthography
use of emphasis	check for added emphasis by capitalizing first letter of a word that is not normally capitalized	orthography
use of caps	check if whole words are written in capital letters	orthography
use of non-standard spelling	check for use of alternative forms, such as wanna, gonna, gotta, kinda in tweet	phonology/orthography
omission of graphemes	check for omission of graphemes to indicate dialectal variety, in particular dropping the h at the beginning of words (ere instead of here)	phonology/orthography
g-dropping	check if final g dropping is evident, e.g., givin', havin", everythin' (with or without apostrophe)	phonology/orthography
iteration of punctuation	check if repeated punctuation (!, ?, or .) is used	orthography/punctuation
use of tweetfinal full stops	check if tweet ends with a full stop	punctuation
use of alternative forms of you	check for use of yous or youse as an alternative plural form of you	lexis
use of swear words	check for use of swear words, e.g., fuck, bloody, f***, fucking, shit	lexis
use of acronyms / abbreviations	check if tweet contains one or more acronyms and/or abbreviations	lexis/morphology
use of contractions	check for use of reduced verbs (forms of to be, to have, e.g., I'm, you're, you've with or without apostrophe)	morphosyntax
use of interjections	use of interjections, e.g., Eurgh, Oh my god, and response particles, e.g., yes, yep, yeah, no, nope	lexis/pragmatics
use of emoticons	check for use of emoticons, e.g. :-)	pragmatics
omission of subject pronoun	omission of subject pronouns - diary drop (e.g., went to the supermarket, bought some soda)	syntax
coordination	check if tweet contains coordinated sentence structure (X is a wonderful product and most people think that it is amazing.)	syntax
no. of words in tweet	count number of words (characters) in tweet	(computational measure)

we use the Yahoo! Placemaker API[5] to disambiguate location information, e.g., "perth wa" is resolved to "Perth (AU)". Finally, we normalized the resulting locations to the country level by identifying the corresponding ISO 3166 country codes.

Name. We used baby names to filter users within their respective countries. Our intuition is that baby names should help remove company names, Twitter bots, etc. and, more importantly, they can serve as a good approximation to sample locals (and hence native speakers). The sources are provided in Table 3.

US: The names from social security card applications[6] include both state-specific data and national data based on social security records as of March 2, 2014 for the year of birth starting with 1910. We gathered a total of 92,599 unique names.

GB: We used baby names for England and Wales from 1996-2013. From this we compiled a list of 29,875 distinct names.

AU: Baby names were compiled from the offices for births, deaths, and marriages in Australia for South Australia, Queensland, New South Wales, Western Australia, the North Territory, Victoria, and Tasmania.

We then removed users whose first name was not contained in the list corresponding to the country we identified based on their location information as described before.

Language. We removed all users who have not set the *language* field in their Twitter profile to 'English'.

Activity. We removed all users who had less than 10 tweets at the time of our crawl.

Overall, from the initial 449,589 users 288,671 (64.20%) were removed.

4.2.2 Tweet Level Filtering

Retweet filter..

Since retweeting other tweets is nothing more than quoting what others say, and mostly without any modification, in our study we only consider original tweets by users and not retweets. We identify retweets by checking if a tweet begins with *RT @username* or *via @username*.

Word Count.

We removed tweets containing less than a threshold of two words.

Language.

Twitter assigns a language to each tweet. However, it does not distinguish between different varieties. We keep only tweets that are identified by Twitter to be written in English, i.e., having been assigned the language flag 'en'.

4.3 Expert Evaluation

To better understand the difficulty of the task and put the results of the classifiers in perspective, we prepared an expert evaluation tasks both at tweet level and user level categorization of English varieties on Twitter. The plain text representation of the tweets was presented to the experts to restrict them from using profile information to infer variety. They were instructed to use their domain

[5] http://developer.yahoo.com/boss/geo/

[6] http://www.ssa.gov/oact/babynames/background.html

Table 3: The data sources for the baby names used to filter users.

	dataset	URL
US	state-specific	http://www.ssa.gov/oact/babynames/state/namesbystate.zip
US	national data	http://www.ssa.gov/oact/babynames/names.zip
GB	England and Wales	http://data.gov.uk/dataset/baby_names_england_and_wales
AU	South Australia	https://data.sa.gov.au/dataset/9849aa7f-e316-426e-8ab5-74658a62c7e6
AU	Queensland	https://data.qld.gov.au/dataset/2010-top-100-baby-names
AU	New South Wales	http://www.bdm.nsw.gov.au/Pages/about-us/facts-statistics.aspx
AU	Western Australia	http://www.bdm.dotag.wa.gov.au/_apps/babynames/RankedBabyNames.aspx
AU	North Territory	http://www.nt.gov.au/justice/bdm/popnames.shtml
AU	Victoria	https://online.justice.vic.gov.au/bdm/popular-names
AU	Tasmania	http://www.justice.tas.gov.au/bdm/top_baby_names

Table 4: Tweet level inter annotator agreement

	US	GB	AU	NA
US	**25**	8	6	2
GB	7	**25**	3	5
AU	0	2	**6**	0
NA	20	15	11	**15**

Table 5: User level inter annotator agreement

	US	GB	AU	NA
US	**40**	8	9	0
GB	1	**39**	7	0
AU	3	1	**32**	0
NA	5	2	2	**1**

knowledge, web search etc., but not try to access the posts on Twitter. The choices for both tasks are the same. Experts were asked to choose the appropriate category of a tweet (or set of tweets) from US, GB, AU or NA if they were not sure.

4.4 Training the Classifiers

4.4.1 Tweet Level

Having generated different sets of features, we applied supervised machine learning algorithms to learn models for classifying the language varieties of tweets. Training and testing sets were generated with stratified sampling by selecting 5,000 tweets from each city. The models for the Linear SVM and Naive Bayes algorithms were learned using 5-fold cross-validation with the scikit-learn[7] library.

4.4.2 User Level

We perform user-level variety identification by regarding up to 20 sample tweets of a user as one document. The rationale is that usually we are interested to study which variety of English a user speaks to perform further analyses at the user level. Furthermore, comparable approaches in the literature, e.g., [7], also work at the user level. Therefore, this experiment helps us to relate our results to prior research in this field.

5. RESULTS AND DISCUSSION

5.1 Expert Evaluation

The contingency tables for the tweet and user level annotation tasks are shown in Table 4 and Table 5. The Cohen's Kappa for the two raters is 0.30 and 0.63 for the two tasks respectively. This shows that both tasks are generally hard but the tweet level categorization is much harder. The balanced accuracy per variety is 0.71, 0.67, 0.57 for the tweet level and 0.87, 0.87 and 0.83 for the user level annotation for US, GB and AU respectively. We observed that US and GB are comparable and AU is more difficult in both tasks.

5.2 Classification Results

Table 6 shows the classification results for the tweet level classification. We see that character Ngrams outperform the linguistic features and the result is acceptable in light of the expert validation.

Table 6: Tweet level classification

feature	algorithm	precision	recall	F1
linguistic	Random Forest	0.36	0.37	0.36
bag of words	Naive Bayes	0.52	0.52	0.52
char Ngrams	**Naive Bayes**	**0.64**	**0.64**	**0.63**

Similarly, in Table 7 we see again character Ngrams outperform the linguistic features at the user level variety classification task.

Table 7: User level classification

feature	algorithm	precision	recall	F1
linguistic	Linear SVM	0.46	0.47	0.46
bag of words	Naive Bayes	0.75	0.75	0.75
char Ngrams	**Random Forest**	**0.79**	**0.77**	**0.77**

6. CONCLUSION AND FUTURE WORK

In this work, we have shown how to build a ground-truth dataset for identifying language varieties on Twitter. We examined the difficulty of the task using expert evaluation. Finally, we experimented with automatic models to perform the task. We found that character Ngrams though simple, are quite effective to classify English tweets by variety.

In future work, we would like to explore automatic identification of more varieties of English as well as other language varieties. We also plan to compare more recent developments in deep learning against linguistic and statistical features for language variety detection.

7. REFERENCES

[1] D. Bamman, J. Eisenstein, and T. Schnoebelen. Gender in twitter: Styles, stances, and social networks. *arXiv preprint arXiv:1210.4567*, 2012.

[2] K. Bolton. World englishes and linguistic landscapes. *World Englishes*, 31(1):30–33, 2012.

[3] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[7] http://scikit-learn.org/

[4] S. C. Herring. Computer-mediated discourse. In D. Schiffrin, D. Tannen, and H. Hamilton, editors, *The Handbook of Discourse Analysis*, pages 612–634. Blackwell Publishers, Oxford, 2001.

[5] J. Jenkins. *World Englishes: A resource book for students*. Psychology Press, 2003.

[6] S. Kersten and N. Lotze. Microblogs global: Englisch. In T. Siever and P. Schlobinski, editors, *Microblogs global. Eine internationale Studie zu Twitter & Co. aus der Perspektive von zehn Sprachen und elf Ländern*, pages 75 – 112. Frankfurt/M, 2013.

[7] M. Lui and P. Cook. Classifying english documents by national dialect. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 5–15, 2013.

[8] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *HLT-NAACL*, pages 380–390, 2013.

[9] M. Purver. A simple baseline for discriminating similar languages. *COLING 2014*, page 155, 2014.

[10] T. Siever and P. Schlobinski, editors. *Microblogs global. Eine internationale Studie zu Twitter & Co. aus der Perspektive von zehn Sprachen und elf Ländern*, volume 4 of *Sprache – Medien – Innovationen*. Peter Lang GmbH, Internationaler Verlag der Wissenschaften, Frankfurt/Main, 2013.